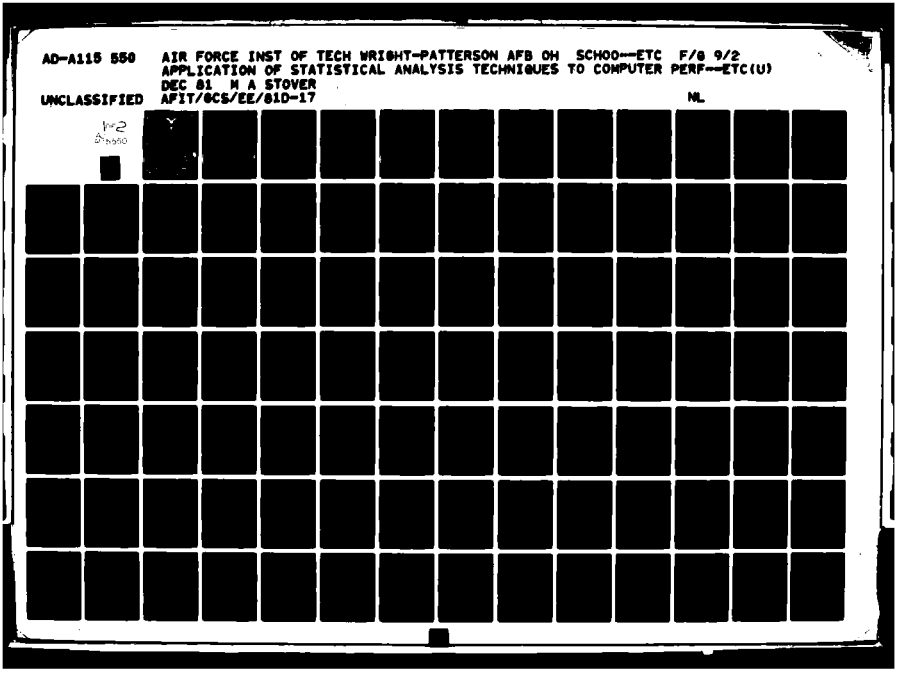
AD-A115 550 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 9/2
APPLICATION OF STATISTICAL ANALYSIS TECHNIQUES TO COMPUTER PERF--ETC(U)
DEC 81 M A STOVER

UNCLASSIFIED AFIT/GCS/EE/81D-17 NL



DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY (ATC)

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GCS/EE/81D-17

APPLICATION OF STATISTICAL ANALYSIS TECHNIQUES TO COMPUTER PERFORMANCE EVALUATION

THESIS

AFIT/GCS/EE/81D-17 Margaret A. Stover
Capt USAF

AFIT/GCS/EE/81D-17

APPLICATION OF STATISTICAL ANALYSIS TECHNIQUES TO COMPUTER PERFORMANCE EVALUATION

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

by

Margaret A. Stover, B.A.

Capt USAF

Graduate Computer Sciences

December 1981

Preface

This project was designed to study the applicability of several multivariate statistical analysis techniques to computer performance evaluation. Such techniques are needed to better explain variable interaction and the complexities of current systems, and to predict future performance of such systems. Specific techniques examined in this effort were multilinear regression, automatic interaction detection, and ridge regression. The techniques were applied to simulated data created through the computer performance evaluation simulation developed at the Air Force Institute of Technology.

I would like to thank my advisors, Dr. Thomas C. Hartrum, Lt Col James N. Bexfield, and Lt Col Charles W. McNichols of the Air Force Institute of Technology for their timely advice and guidance concerning this study. I would also like to thank Capt Patricia K. Lawlis of the Air Force Institute of Technology for the many hours she made her personal computer available for this project.

Margaret A. Stover

TABLE OF CONTENTS

		Page
Preface		ii
List of Figures		v
List of Tables		vi
Abstract		viii
I.	Introduction	1
	Problem	3
	Scope	4
	Assumptions	6
	Summary of Current Knowledge	6
	Approach	7
	Materials and Equipment	8
	Expected Results	8
II.	Initial Data Analysis	9
	Workload Characterization	9
	CPESIM Data Description	12
	Statistical Description	12
	Multilinear Regression	16
	CPESIM Regression Results	19
III.	Automatic Interaction Detection (AID)	31
	Theoretical Discussion	31
	CPESIM Data Processing	36
	Turnaround Time Model	48
	IO Time Model	60
IV.	Ridge Regression	72
	Theoretical Discussion	72
	CPESIM Data Processing	76
	Turnaround Time Model	76
	IO Time Model	85
V.	Comparison of Techniques	94
	Explained Criterion Variability	94
	Results Available	94
	Ease of Use	97
	Algorithm Limitations	99
	CPESIM Summary	99

Page

VI. Conclusions and Recommendations 100

Bibliography 102

Vita 105

List of Figures

Figure		Page
1	Scattergram of Turnaround Time (Down) with Tapes (Across) . .	28
2	Scattergram of Turnaround Time (Down) with Memory (Across) .	29
3	Scattergram of IO Time (Down) with TapeIO (Across).	30
4	Sample AID Detailed Tree	45
5	AID Turnaround Time Detailed Tree	46
6	AID IO Time Detailed Tree	47
7	Turnaround Time Ridge Trace	81
8	IO Time Ridge Trace	90

List of Tables

Table		Page
I	Predictor and Criterion Variables Included in the Model . .	13
II	Workload Parameter Statistics Initial Data File	14
III	Workload Parameter Statistics Test Data File	15
IV	Turnaround Time Regression Factors Initial Data File . . .	20
V	Turnaround Time Regression Factors Test Data File	22
VI	IO Time Regression Factors Initial Data File	23
VII	IO Time Regression Factors Test Data File	25
VIII	Multilinear Regression Equations Evaluation Statistics . .	26
IX	Variable Length AID Intervals	38
X	AID Sample Detailed Output	43
XI	AID Sample Summary Output	44
XII	AID Summary Division Data Turnaround Time Sequential Splits	49
XIII	AID Detailed Tree Data Turnaround Time	51
XIV	Workload Parameters Statistics Turnaround Time - Group 4 .	53
XV	Workload Parameter Statistics Turnaround Time - Group 5 . .	54
XVI	Workload Parameter Statistics Turnaround Time - Group 6 . .	55
XVII	Workload Parameter Statistics Turnaround Time - Group 7 . .	56
XVIII	Turnaround Time T-Test Comparisons Between Initial and Test Data Files	58
XIX	Turnaround Time T-Test Comparisons Within Initial Data Files	59
XX	AID Summary Division Data IO Time Sequential Splits	61
XXI	AID Detailed Tree Data IO Time	63
XXII	Workload Parameter Statistics IO Time - Group 4	65
XXIII	Workload Parameter Statistics IO Time - Group 5	66

Table		Page
XXIV	Workload Parameter Statistics IO Time - Group 6	67
XXV	Workload Parameter Statistics IO Time - Group 7	68
XXVI	IO Time T-Test Comparisons Between Initial and Test Data Files	69
XXVII	IO Time T-Test Comparisons Within Initial Data File	71
XXVIII	Turnaround Time Correlation Matrix	77
XXIX	Turnaround Time Normalized Regression Coefficients	79
XXX	Turnaround Time Unnormalized Regression Coefficients . . .	80
XXXI	Turnaround Time Variance Inflation Factors	82
XXXII	Turnaround Time Ridge Regression Evaluation Statistics . .	84
XXXIII	IO Time Correlation Matrix	86
XXXIV	IO Time Normalized Regression Coefficients	87
XXXV	IO Time Unnormalized Regression Coefficients	88
XXXVI	IO Time Variance Inflation Factors	91
XXXVII	IO Time Ridge Regression Evaluation Statistics	92
XXXVIII	Accuracy Comparison of Multilinear and Ridge Regressions .	96

Abstract

Multivariate statistical analysis techniques were examined to determine their applicability to computer performance evaluation. Specific techniques were multilinear regression, automatic interaction detection (AID), and ridge regression. A simulated data base created through a computer performance evaluation simulation developed at the Air Force Institute of Technology was modeled using each technique. Multilinear regression derived the most accurate modeling equations, though no method adequately explained the variation in turnaround time. Ridge regression, designed to circumvent multicollinearity, also created modeling equations, but was less accurate because no multicollinearity was present in the system. AID, however, clustered the data into groups which explained high levels of the variability of the criterion variable within each group, and little between the groups. Data records which met the selection criteria for each group should also have values of the criterion variable within the range specified by that group. Although difficult to use, AID provided initial information about systems which could not be modeled through regression or other techniques. The applicability of each technique to selected types of computer systems was also documented.

APPLICATION OF STATISTICAL ANALYSIS TECHNIQUES TO COMPUTER PERFORMANCE EVALUATION

I. Introduction

Computer systems are designed to process information quickly and with the most efficient use of system resources. Therefore, computer systems are evaluated on how well they perform required processing functions. Three major performance criteria of a computer system are total system throughput (the number of jobs processed in a specified period), the turnaround time of individual batch programs, and remote response time in a real-time oriented environment. A multiprogramming environment increases throughput by processing several computer programs simultaneously. However, individual program turnaround times and remote user response times could actually decrease because each program now shares and competes for system resources. Therefore, evaluation and control of computer workload and resources are vital to insure proper computer support (Ref 8:2).

Computer performance evaluation (CPE) assesses a system's performance to produce an overall understanding of how well the system performs its various functions. CPE activities involve collecting, analyzing, and interpreting computer performance data for both workload support and resource utilization. The analyst obtains the required data either directly from the system or its representation, and models the system using analytic, simulation, or empirical techniques. Measurement techniques study the current system using actual data collected directly from the operational system. System models are organized representations of the system functions built to study the system by relating dependent (performance) variables to independent (workload) variables.

Analytic models present the system as a series of equations relating system parameters to performance criteria. These equations, representing a discrete set of parameters, are then mathematically solved. This may be an iterative process to calculate the values of the dependent variables from several combinations of the independent variables (Refs 17:17; and 9-45-7). Thus analytic models evaluate system performance with minimal costs and effort for a variety of system parameters and configurations. Further, the effects of various changes to the system configuration and workload can be evaluated before any are actually made. A major analytic method is queuing theory which views a computer system as a multiple resource system and models contention for the resources. While valuable, and frequently quite accruate in its prediction results, queuing theory is constrained when working with complex multiple resource systems. Further, the analytic model developed may be impossible to solve mathematically. Analytic models are also criticized for too often making too many simplifying assumptions about the system, and thus biasing the model (Ref 24:11).

Simulation is another numerical modeling technique, frequently performed on a digital computer. A simulation recreates the system performance over time, studying the evolution and dynamic behavior of the system. The parameters of interest are tracked as they occur chronologically, working under or mimicking an environment identical to the actual system. In addition, a simulation may model a system which does not yet exist. Further, for many simulations, the workload parameters could be easily adjusted to emulate different environment changes. Here agin, the analyst may study the effects of these changes without actually making them. Difficulties with this technique are building the simulation and insuring its correctness (Refs 24:11; and 17:18).

Empirical models characterize the computer system performance statistically, based on either real or simulated data. The interactions of functional and performance variables are often too complex to be readily understandable when the logical system structure is considered alone. Because empirical models are based totally on experimental data, they go beyond the logical system to its results. Besides modeling the system, these measurement tools often detect system design or implementation deficiencies by revealing unexpected system behavior under select conditions. Further, the statistical techniques are also often capable of identifying the degree of confidence (or uncertainty) associated with the model (Ref 17:21-22).

These modeling techniques characterize and precisely describe the workload of a system to understand what resource needs correspond to different workload elements. Therefore, the techniques explain the performance of current systems, predict future performance of existing or proposed systems, or verify improvements to working systems. Appropriately designed models can also be used to tune existing systems to improve system efficiency and user satisfaction (Ref 18:4). Various analytic and simulation modeling techniques have been developed and widely used for CPE, but have not proven completely satisfactory for all systems. Thus, this thesis effort studied the less explored area of empirical modeling techniques.

Problem

Major drawbacks exist with the statistical techniques most commonly used to model systems for CPE. The techniques frequently assume the dependent variables are linearly related to the independent variables, or attempt to transform the data to impose linearity before processing the data. In addition, some techniques assume no interaction exists among the independent

variables, a condition called nonadditivity. Finally, the independent variables are frequently deemed to be truly uncorrelated, i.e., no multicollinearity exists. However, data are frequently nonlinear, effective data transform algorithms are difficult to develop and apply, and variables in a system with multiple workload parameters are rarely independent. Thus current techniques, while valuable, have limitations and newer multivariate modeling performance techniques should be considered.

The techniques attempt to account for nonlinear variables and their interrelationships. These techniques also consider multiple performance parameters, and parameters whose influences could not be considered individually. Further, they study the performance of subsets or clusters of job or activity types where data nonlinearity cannot be otherwise managed.

Scope

This thesis effort examined several multivariate analysis techniques to determine their applicability to CPE. These statistical techniques were designed to deal with large-size multidimensional samples. Thus, they accounted for multiple factors in a typical computer system workload analysis. No effort was planned or made to develop totally new techniques; only existing methods were considered. Multilinear regression was first studied to provide insight into the system being studied and to establish a baseline against which to compare results of the other techniques tested. The multivariate techniques considered on a theoretical basis for possible experimentation included: Factor analysis, canonical correlation analysis, cluster analysis, discriminant analysis, automatic interaction detection (AID), and ridge regression. The first four methods are among the more commonly used multivariate techniques capable of handling the multiplicity

of variables involved in CPE. The latter two techniques were included on the recommendations of the thesis committee.

Factor analysis. This quantitative method collects and analyzes data on a set of variables to determine the variable interdependence or structure. It isolates the major statistical dimensions or factors in the data sample. Thus it reduces a large number of variables, often highly correlated, into a smaller number of underlying factors, while simultaneously reducing variable redundancy (Ref 21:6-1-6-5).

Canonical correlation analysis. This technique initially defines sets of predictor and criterion variable observations. It then attempts to establish a linear combination of the performance variables and a second linear combination of the criterion variables in such a fashion as to minimize the correlation between the two linear combinations (Ref 21-5-1-5-4).

Cluster analysis. This method partitions the data set into homogeneous and well-separated subsets. The goal is to minimize the maximum dissimilarity between entries in a single cluster while maximizing the dissimilarities between the clusters themselves (Ref 29:147-8).

Discriminant analysis. This method also assigns an individual record from the data set to one of several distinct populations by observing characteristics of the data. It attempts to determine linear combinations of the original independent variables which show the largest differences in the group means (Ref 23:435-6).

Automatic interaction detection (AID). This technique is also similar to clustering in that it iteratively divides the data set into smaller groupings with as much dissimilarity between or among the groupings as possible. The group with the greatest variability remaining is selected for each division (Ref 21:8-1-8-4).

Ridge regression. This technique uses an alternative method (to the least squares method used by multilinear regression) to determine the most significant variables affecting the performance variables. It is particularly useful in explaining interrelations between a criterion variable and highly nonorthogonal predictor variables (Ref 3:181).

Not all the techniques considered theoretically were put to a practical test. Ridge regression and AID were selected due to the information available and the lack of any published reports of CPE studies using them.

Assumptions

Two assumptions were made while designing this thesis project.

1. The first was that the selected techniques would be available for actual testing on either the Cyber or Create computer systems. Factor analysis, discriminant analysis, and canonical correlation analysis were available through the Statistical Package for the Social Sciences (SPSS). Independent modules were locally available for AID, cluster analysis, and ridge regression. All were supported on the Cyber computer system.

2. The second assumption was that actual or simulated CPE data would be available to test the selected techniques. Simulated system accounting data were available from the CPE simulation developed at the Air Force Institute of Technology for EE752, Computer Performance Evaluation (Ref 9). The MAC Washington Area Computer Center, Det 1, 1500 Computer Services Squadron, Andrews AFB, Md., provided actual monthly workload accounting data from its dual Burroughs B4700 configuration, though these data were not used due to time constraints.

Summary of Current Knowledge

Although some initial work has been done applying multivariate analysis

techniques to CPE, their use is by no means universal. A. K. Agrawala and others (Refs 1; and 29) have applied cluster analysis, though more to modeling computer workload than computer performance. In addition, some attention has been paid to factor analysis by D. R. Cruise among others (Refs 4; and 22) and to discriminant analysis by Little (Ref 19). Although these techniques were not tested during this effort, the work which had been done using them also provided a base from which to evaluate the results of this thesis.

Approach

Precisely defining this thesis problem was initially difficult. The exact techniques to be tested could not be specified immediately, since no work had occurred to determine which of them were most feasible. The first consideration was to determine what techniques were currently available and which might apply to CPE. Therefore, a literature search of multivariate analysis techniques and CPE was conducted through statistical and computer-oriented literature. Specific techniques considered, as listed above, were ridge regression, factor analysis, cluster analysis, canonical correlation analysis, discriminant analysis, and AID. Each was studied in order to obtain a detailed statistical understanding to determine the most promising tools for CPE. CPE itself, the concept of performance evaluation, the most important performance indices, and CPE applications were also studied. Multilinear regression was also considered for clarity and insight, providing a basis for later comparisons and evaluations of the technique results. The results of this theoretical mathematical study were applied to practical exercises processing simulated CPE data through the selected techniques. Finally, the results received from the different techniques were analyzed and compared to evaluate the applicability of each technique to CPE.

Materials and Equipment

No special computer hardware was needed to support this project. Software support for the selected techniques was required and was available through SPSS and other independent systems. In addition, both simulated and actual data were available.

Expected Results

This thesis effort determined which of the multivariate analysis techniques studied were most applicable to CPE. Analyzing and comparing the experimental results also produced a hierarchy of applicable tools for the computer system manager.

II. Initial Data Analysis

Workload Characterization

Although the objectives of a given CPE effort vary with the specific problem being studied, the basic areas considered are generally throughput rate, turnaround/response time, and resource use. Models are normally developed to understand the current system, as well as to assess the effects of changes to the workload and/or environment before such changes are made. However, to do this successfully, the analyst must fully understand the interplay between the computer system hardware/software environment and the workload. Hence, collecting and analyzing measurement data from the system under study is vital. These data are collected from the system or its model while a typical workload is being processed. Thus information about the nature and resource requirements of each job is available to form statistical samples of the workload. System accounting logs, hardware monitors, and software monitors are prime sources of representative data used to identify and characterize aspects of system performance (Refs 24:8; 27:150; and 18:4).

The test workload, as well as the measurements taken, determine the results of the experiment. Therefore, care must be taken to insure the representativeness of the test data. To this end, the time interval over which the data are measured must also be representative of the system, and the underlying system characteristics must remain constant, or stationary (Refs 2:1; and 16:5). Although this is virtually impossible to guarantee in practice, the experiment must be designed to provide reasonable assurance that these conditions will be met.

The analyst must understand the system workload before he/she determines the representativeness of the test data. The form of this workload is

directly related to the specific techniques being used in the study. The workload measured and models developed in this thesis are statistical in nature, thus enabling statistical validation of the models. Kelly (Ref 16:1) lists six advantages provided by formal statistical techniques.

1. The data analysis is now completely objective; all subjectivity has been removed.

2. Generalizations about the system are better supported by the statistical information produced.

3. The number of false starts toward creating the model are reduced.

4. Comparing results of different workloads is simpler.

5. Credibility of conclusions drawn is improved.

6. Statistics provide a more scientific framework for the CPE study under development.

However, to use these methods successfully, the analyst must be thoroughly familiar with the system involved. He/she must understand statistical theory to some degree to insure an adequate relationship between the data and the underlying assumptions made. Further, he/she should be familiar with various statistical techniques, their capabilities and limitations, and what information can be derived from each, to insure selection of the desired information. Finally, the analyst must be able to organize sizable masses of data and to present them in logical and readable form (Ref 16:1-2).

Lewis (Ref 18:15-16) identifies five significant characteristics for measurement data.

1. Data is collected in often enormous amounts, up to and including millions of observations, depending on the model and system involved.

2. The relative "times-between-event" frequently includes discrete

components rather than occurring in strictly exponential distributions.

3. Data stationarity cannot always be assumed or proven.

4. Significant differences or irregularities in the data may occur frequently due to the external system variables.

5. The data may suggest a lack of stochastic regularity, or insufficient time or data may be present to determine such a pattern.

These characteristics clearly show the need to fully understand, explain, and document the system workload.

The workload measure must document the actual system load in measurable units, such as CPU time, blocks of memory required/used, etc. These variables are used to establish the load on the individual system resources. Further, the measure must account for the temporary level of workload which varies across time and is directly dependent on the rate of arrival of user jobs. Finally, the workload measure should represent the external, user-oriented characteristics of the user jobs (Ref 7:2). This last requirement was not considered in the CPE simulation (CPESIM) data. The main system resources tracked by CPESIM are documented later.

The analyst first measures the independent and dependent variables across a given time period to establish a baseline and create a model. The model is then validated by measuring and collecting another independent data set and using the model to calculate the values of the dependent variables based on those of the independent variables. The calculated figures are then compared to the actual measured results to validate the accuracy and representativeness of the model (Refs 9:45; and 24:16). Alternatively, the basic model may be independently validated by generating models from separate data bases and comparing the results to determine the stability and consistency of the models (Ref 2:28).

CPESIM Data Description

Data for analysis were collected from simulated accounting log data produced by CPESIM. A five day period was used for the initial data base; six additional days were measured to produce the independent test workload. The variables collected and used in the following models appear in Table I. The assignment of predictor (workload) and/or criterion (performance) status for each variable is also indicated. Although other data parameters were available, they were not deemed suitable for this effort.

The data were merged into a single file for each set to create as representative a workload as possible. The first file contained 787 data records; the second contained 922 records. The system workload was then characterized for initial analysis. All statistical analyses were performed using programs available through SPSS (Ref 23). The stationarity of the workload was studied from day-to-day and from hour-to-hour over a single day. The T-tests showed the workload was stationary for all relevant workload parameters both day-to-day and hour-to-hour. Thus the data sample used was representative of the system operation.

Statistical Description

This thesis effort first examined the workload using descriptive and inferential techniques. Cumulative frequencies (produced by SPSS procedure Condescriptive) were considered for the selected variables to analyze the overall system workload. Table II contains the range, mean, standard deviation, and 95 percent confidence interval for each workload parameter for the first data file. Table III contains the same information for the test data file. The mean of each variable would be expected to fall within the confidence interval 95 percent of the time for any given period (of

Table I

Predictor and Criterion Variables
Included in the Model

Variable	Descriptive Title	Status
Cards	Number of Cards Read	Predictor*
CPUtime	CPU Time in Seconds	Predictor
DiskIO	Number of Disk IOs	Predictor*
IOtime	IO Time in Hours	Predictor/ Criterion
Lines	Number of Lines Printed	Predictor*
Memory	Memory in Blocks	Predictor
TapeIO	Tape Drives Assigned	Predictor*
Turntime	Turnaround Time in Hours	Criterion

*Predictor variables for both turnaround time and IO time

Table II

Workload Parameter Statistics
Initial Data File

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 3146.0	238.968	330.958	215.810 - 262.126
CPUtime	0.0 - 999.0	154.574	201.310	140.488 - 168.661
DiskIO	0.0 - 9999.0	895.422	1237.846	808.806 - 962.038
IOtime	0.0 - 3100.0	302.192	387.660	275.066 - 329.317
Lines	0.0 - 9090.0	1047.529	1398.204	949.692 - 1145.365
Memory	10.0 - 205.0	69.269	51.004	65.701 - 72.838
TapeIO	0.0 - 2628.0	233.305	361.792	207.989 - 258.621
Tapes	0.0 - 3.0	1.409	1.146	1.329 - 1.489
Turntime	0.00(- 5.728	1.549	1.489	1.445 - 1.635

Table III

Workload Parameter Statistics
Test Data File

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 2542.0	263.428	349.106	240.865 - 285.992
CPUtime	0.0 - 999.0	165.467	210.637	151.853 - 179.082
DiskIO	0.0 - 8942.0	917.645	1227.317	838.320 - 996.970
IOtime	0.0 - 2700.0	302.352	383.453	277.578 - 327.145
Lines	0.0 - 10506.0	996.850	1396.974	906.560 - 1087.141
Memory	3.0 - 204.0	66.206	48.760	63.055 - 69.358
TapeIO	0.0 - 2612.0	232.714	355.473	209.738 - 255.689
Tapes	0.0 - 3.0	1.423	1.094	1.352 - 1.494
Turntime	0.009 - 7.003	1.679	1.638	1.573 - 1.785

sufficient length).

Studied also was actual memory allocated across time. This test found a clear grouping of job sizes. The groupings, and the percent of workload in each were approximately:

Memory	Percentages
4K - 36K	36.96%
50K - 68K	36.33%
100K - 150K	20.80%
185K - 205K	5.91%

No jobs required an amount of memory blocks falling between the indicated groups.

Multilinear Regression

Regression analysis determines the statistical form of the relationship between the performance and workload variables, thus quantifying that relationship. To determine this relationship, large numbers of dependent and independent variables must be measured. The regression equation fits a linear analytic function to a set of discrete data points. The functional relationship is assumed to be linear and is represented by the equation (Ref 9:20-21):

$$Y = b(0) + \sum_{i} b(i)x(i) \qquad (1)$$

where Y is the criterion variable

x(i) are independent workload and system parameters

b(i) are regression coefficients produced by the analysis

The SPSS multilinear regression procedure was processed in this project using stepwise regression. Here the program regressed against the performance variable using the single "best" workload parameter. Next, it regressed with the "best two" parameters, continuing similarly until all the variables were included in the equation. The "best" variable was considered to be the one which explained most of the variation (remaining) in the performance variable.

Hartrum (Ref 9:55) provides several assumptions which must be met before a multilinear regression model can justifiably be built.

1. The random error variables e(i) are independent of each workload variable.

2. The e(i) are normally distributed and have a mean of zero for the given set(s) of values.

3. Any two errors, e(i) and e(j), are independent of each other.

4. The variance of each e(i) is constant for any combination of the independent values.

5. No independent variable can be expressed as a simple linear combination of the remaining independent variables.

6. The number of data records collected must be at least equal to the total number of coefficients in the equation.

The most common method used to estimate the regression coefficients, least squares estimation, may be calculated when these conditions are met. The procedure determines those coefficient estimates which minimize the sum-of-squares differences between the predicted value of the performance variable and its actual measured value. The random error, e(i), the unexplained portion of the variable, is the difference between the two

resulting values. Thus least squares estimation seeks to minimize the estimation errors (Ref 21:4-14-4-15).

Two other conditions which may exist with the independent variables could cause inaccuracies in interpreting regression results. The first is multicollinearity which occurs when two or more nonlinearly related variables are nonetheless highly correlated (linearly dependent). The variables must be nonlinearly related to meet condition five. This may significantly impact the regression results because minute changes in one variable involved may result in large changes in the regression coefficients of the other(s) (Refs 9:55; and 21:4-63-4-64). In addition, the model assumes the independent variables are additive, that no interaction exists between them. If two variables x(i) and x(j) are nonadditive, changes in x(i) would cause changes in the coefficient b(j), though not in the variable x(j) itself. Measures to detect and correct nonadditivity are discussed by Hartrum (Ref 9:60-61).

Multilinear regression produces many statistics explaining the resulting equation. Those considered here included the regression coefficients, and the ratio of the unexplained portion of the regression equation to the explained portion, denoted R-squared. This coefficient of determination is a reasonable index of how well the data fit the computed linear regression model. When stepwise regression is used, changes in the F-number, its corresponding level of significance, and the overall-F-number with its level of significance should also be studied. The first two statistics document whether the change in R-squared is statistically significant or merely due to variation in the data. The overall-F-number reflects the degree to which the most recently added variable contributes to the explanation of the model. If it adds little, this value may decrease. A rule of

thumb for using the overall-F-number states once this value starts to decrease, the analyst may often consider the remaining variables as adding too little explanation onto the model to be included in the equation (Refs 21:4-24-4-25; and 9:59).

Two cautions should be considered when performing regression analysis. First, while regression always produces coefficients for a regression equation, the analyst should not assume the equation itself guarantees a direct cause and effect relationship. Other factors not considered or perhaps not measurable may affect the performance criterion. Nevertheless, the model may serve adequately to better understand the system or predict future performance. Secondly, if the criterion variable ranges beyond the workload variables, the analyst should not assume the within-range results hold true automatically outside the range. They may, in fact, but this would depend on the system and workload being measured and could not be guaranteed (Ref 16:7).

CPESIM Regression Results

Two performance variables were studied, turnaround time and IO time. The regression equation for turnaround time for the first file was:

$$Y = -.079197928 + .71613196*Tapes + .008147729*Memory \qquad (2)$$

This equation explained .50910 percent of the variation in the values of turnaround time. The significant order of the factors involved appears in Table IV. Memory was included when the rule of thumb would have omitted it because it added seven per cent to the explanation and provided a more complete picture of the system. Multicollinearity may have been suspected between IOtime and TapeIO since the regression correlation coefficient,

Table IV

Turnaround Time Regression Factors
Initial Data File

Variable	F-Number	Significance	R-Square	Overall-F	Significance
Tapes	605.98758	.000	.43565	605.98758	.000
Memory	117.30830	.000	.50910	406.54054	.000
IOtime	3.63451	.057	.51137	273.14928	.000
CPUtime	2.57751	.109	.51298	205.91907	.000
Lines	3.31977	.069	.51504	165.88789	.000
TapeIO	1.68726	.194	.51609	138.64277	.000
DiskIO	1.85067	.174	.51723	119.23064	.000
Cards	.75915	.384	.51770	104.38945	.000

.96920, was significantly larger than that of the overall equation, .50910. However, these variables were not included in the regression equation and hence this multicollinearity should not have biased the final model. Further, this suspected multicollinearity was another rationale to inspect the effect of predictor variables on IO time.

The regression equation for IO time for the first data file was:

$$Y = 1.8786988 + .99121244*TapeIO + .067400188*DiskIO \qquad (3)$$

This equation explained .98629 percent of the variation in the value of IO time. The significant order of the factors involved appears in Table V. Both variables included met the rule of thumb guidance. Multicollinearity was not suspected because none of the individual regression correlation coefficients were significantly larger than that of the overall equation.

The regression equation for turnaround time for the test data set was:

$$Y = -.048865649 + .81748272*Tapes + .0047164398*Memory \qquad (4)$$

This equation explained .46975 percent of the variation in turnaround time. The significant order of the factors involved appears in Table VI. Again, multicollinearity could have been suspected between IOtime and TapeIO since the regression correlation coefficient, .97364, was significantly larger than that of the overall equation, .46975. Again, however, these factors were not included in the system model and hence would not have biased it unduly.

Finally, the regression equation for IO time for the test data file was:

$$Y = 1.1586177 + .99518033*TapeIO + .067622933*DiskIO \qquad (5)$$

Table V

Turnaround Time Regression Factors
Test Data File

Variable	F-Number	Significance	R-Square	Overall-F	Significance
Tapes	747.25366	.000	.44819	747.25366	.000
Memory	37.35186	.000	.46975	407.06583	.000
IOtime	19.64198	.000	.48085	283.42945	.000
CPUtime	4.69867	.030	.48350	214.60322	.000
Lines	1.20024	.274	.48418	171.96011	.000
Cards	.91492	.339	.48469	143.43927	.000
TapeIO	.61673	.432	.48504	122.98455	.000
DiskIO	.86973	.351	.48553	107.70486	.000

Table VI

IO Time Regression Factors
Initial Data File

Variable	F-Number	Significance	R-Square	Overall-F	Significance
TapeIO	12159.22188	.000	.93936	12159.22188	.000
DiskIO	2683.23585	.000	.98629	28194.41415	.000
Lines	29.51049	.000	.98679	19489.64736	.000
Cards	2.55819	.110	.98683	14646.96376	.000
Tapes	.01131	.915	.98683	11702.75857	.000

This equation explained .99439 percent of the variation in the values of IO time. The significant order of the factors involved appears in Table VII. Again, the variables included met the rule of thumb guidance. Further, multicollinearity was again not suspected since the regression correlation coefficients were all significantly less than that of the overall equation.

Model. After multilinear regression is performed, the regression equation is normally used to model the system. However, its accuracy needs to be verified. Equation 2 is particularly doubtful, since it explains only .50910 percent of the turnaround time variation. Equation 3 is expected to be more accurate, since it explains fully .98629 percent of the IO time variation. To verify the equations, 14 data records were chosen randomly from the test data file. Equations 2 and 3 were applied using the appropriate variable values. Table VIII contains the predictor values used, the actual and calculated values of the criterion variable, and the percentage error between them for both turnaround time and IO time.

The percent error for turnaround time ranged from -48.71 percent to 1183.62 percent. The average relative error was 155.85 percent, with an average absolute error equaling 182.6 percent. The largest value was an outlier, falling fully 607.81 percentage points (105.56 percent) beyond the next greatest error value. In addition, the actual value was 0.0116 and the estimated value 0.1489, a relative increase of only 0.1373, though a percentage increase of 1183.62 percent. However, even with this one error omitted, the mean relative error was 76.79 percent and the mean absolute error was 105.60 percent. This confirms that the regression equation 2 is inadequate to model turnaround time. Further effort is required to fully explain the variation in turnaround time. This result is also confirmed

Table VII

IO Time Regression Factors
Test Data File

Variable	F-Number	Significance	R-Square	Overall-F	Significance
TapeIO	16762.60156	.000	.94797	16762.60156	.000
DiskIO	7603.50997	.000	.99439	81442.75439	.000
Lines	164.93739	.000	.99524	64035.68571	.000
Cards	2.28641	.131	.99526	48094.63687	.000
Tapes	2.42585	.120	.99527	38536.02089	.000

Table VIII

Multilinear Regression Equations Evaluation Statistics

Memory	Tapes	Turnaround Time Actual	Turnaround Time Calculated	Percent Error	TapeIO	DiskIO	IO Time Actual	IO Time Calculated	Percent Error
14	2	2.3121	1.4671	-36.55	238	226	259.1235	247.0724	-4.65
20	0	0.0124	0.0838	575.81	0	28	1.9877	3.7659	89.46
64	3	1.5448	2.5817	67.12	385	2505	553.9414	552.3330	-0.29
116	2	1.0676	2.2982	115.27	299	1963	454.9939	430.5578	5.37
54	0	0.1243	0.3608	190.27	0	437	34.9514	31.3326	-10.35
28	1	0.9913	0.8651	-12.73	693	534	814.8247	723.3388	-11.23
27	0	0.0429	0.1408	228.20	0	300	22.3143	22.0988	-0.97
14	2	2.1271	1.4671	-31.03	36	2679	210.5129	218.1274	3.62
51	2	3.4483	1.7686	-48.71	181	866	242.3307	244.8626	1.04
149	3	4.3499	3.2743	-24.73	807	37	809.4993	804.2809	-0.64
28	0	0.0116	0.1489	1183.62	0	112	7.8876	9.4275	20.29
32	2	1.5332	1.6138	5.26	323	304	357.2274	342.5300	-4.11
59	3	2.4613	2.5499	3.60	85	248	113.1524	98.6008	-12.86
50	2	2.6472	1.7605	-33.50	31	185	50.8446	45.0753	-11.35

visually in Figures 1 and 2 (produced by SPSS Scattergram procedure) which show the effect on turnaround time of tapes and memory, respectively. The values of turnaround time are not significantly different for any of the tape or memory groups, thus reaffirming the lack of a significant linear relationship between the predictor and criterion variables.

However, the percent error for IO time ranged only from -0.29 to 89.46 percent. The average relative error was 3.76 percent, with the average absolute error equaling 12.59 percent. Again, the largest value was also an outlier, 69.17 percentage points (340.91 percent) greater than the next largest error. Further, this particular job required only 28 diskIOs and no tapeIOs for an equation heavily weighted toward the number of tapeIOs. Omitting this error reduced the mean relative error to -2.84 percent, and the mean absolute error to 6.67 percent. Thus, the validity of using regression equation 3 to predict IO time is confirmed. Figure 3 (again produced by SPSS Scattergram) clearly shows the almost linear relationship between tapeIOs and IO time.

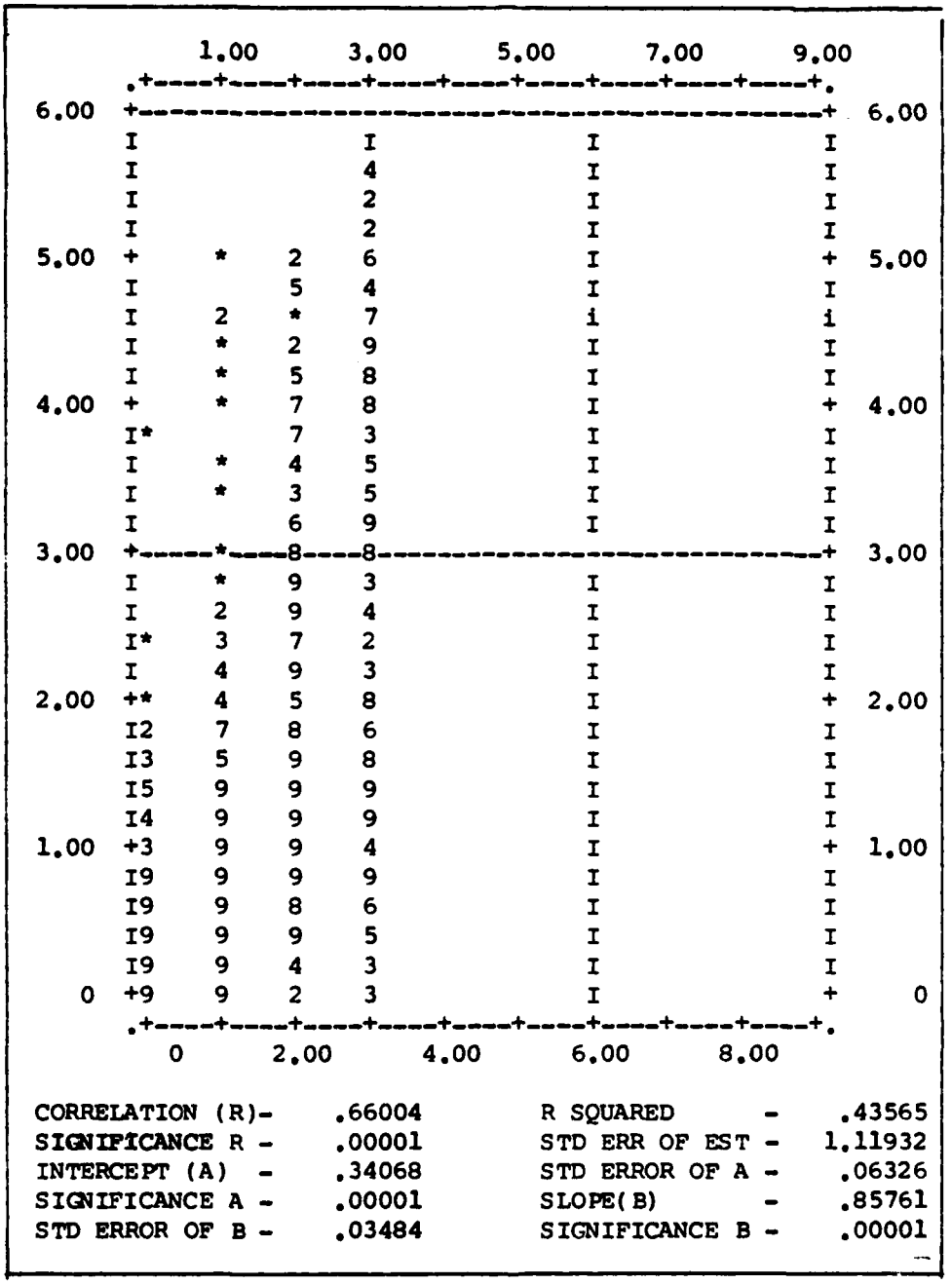
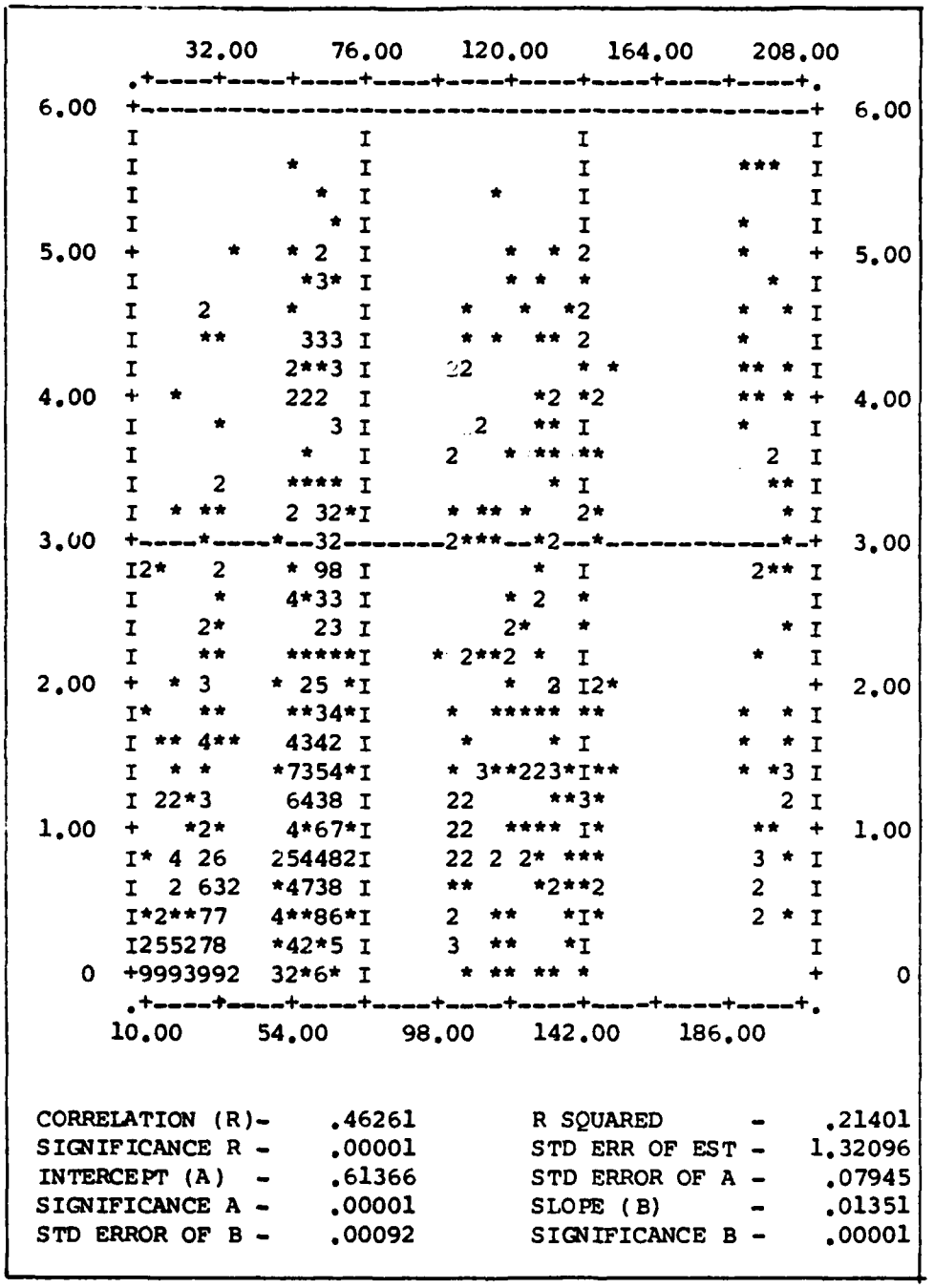
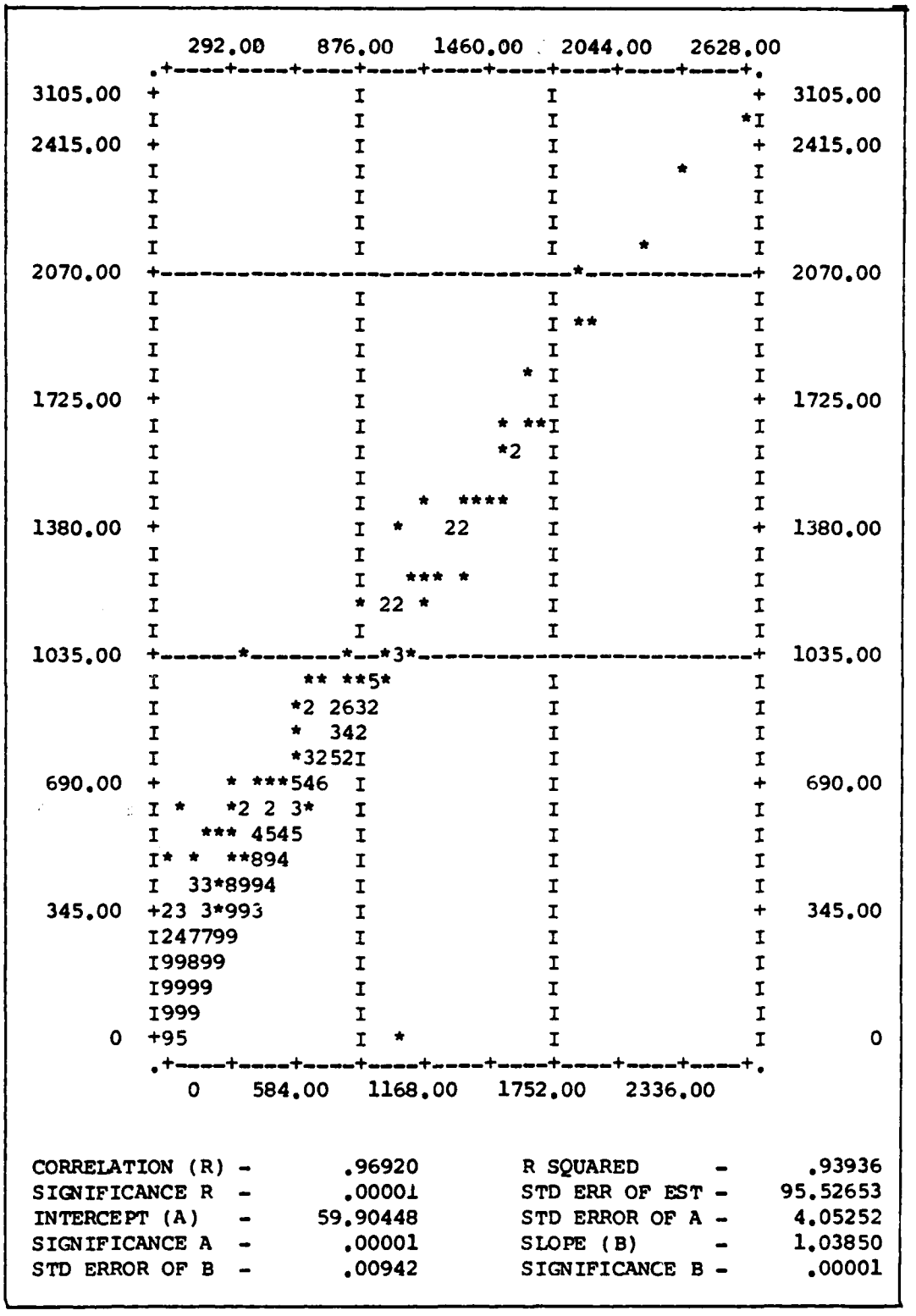


Figure 1. Scattergram of Turnaround Time (Down) with Tapes (Across)



CORRELATION (R)-	.46261	R SQUARED -	.21401
SIGNIFICANCE R -	.00001	STD ERR OF EST -	1.32096
INTERCEPT (A) -	.61366	STD ERROR OF A -	.07945
SIGNIFICANCE A -	.00001	SLOPE (B) -	.01351
STD ERROR OF B -	.00092	SIGNIFICANCE B -	.00001

Figure 2. Scattergram of Turnaround Time (Down) with Memory (Across)



CORRELATION (R) -	.96920	R SQUARED -	.93936
SIGNIFICANCE R -	.00001	STD ERR OF EST -	95.52653
INTERCEPT (A) -	59.90448	STD ERROR OF A -	4.05252
SIGNIFICANCE A -	.00001	SLOPE (B) -	1.03850
STD ERROR OF B -	.00942	SIGNIFICANCE B -	.00001

Figure 3. Scattergram of IO Time (Down) with TapeIO (Across)

III. Automatic Interaction Detection (AID)

Theoretical Discussion

The Automatic Interaction Detection (AID) algorithm constructs models to analyze interdependence between a single criterion variable and several or many predictor variables. The former must be interval scaled; the latter may include interval, ordinal, and/or nominal scaled variables. AID partitions the total data base into mutually exclusive groups defined by similar responses (or ranges of values) for subsets of the predictor variables. Group elements would also have similar values of the criterion variable. These groups can then be examined to analyze the dependence of subsets of the criterion variable on the predictor variables. This algorithm was developed by J. A. Sonquist and J. N. Morgan of the University of Michigan's Institute for Social Research (Refs 6:62; and 21:8-1-8-4).

AID partitions the data base into subsets through a sequence of two-way splits. The entire data set is first partitioned into two subsets in such a way as to minimize the within-group variability, the variance of the criterion variable for elements in a given group, and maximize the between-groups variability, the variance of the criterion variable between selected groups. This minimization is computed by the sum-of-squares deviations of the criterion variable from the group means. The partition selected most greatly reduces the remaining unexplained variability of the criterion variable. For the second iteration, the subgroup with the greater variability is partitioned in a similar manner. The process continues, each time selecting the subgroup with the largest variability for partitioning, until it reaches one of several stopping criteria discussed below (Refs 6:63; and 21:8-4).

Mathematical Development. The algorithm selects the best partition by applying a one-way analysis of variance over each possible split in the predictor variables. The algorithm calculates the mean and variance of the criterion variable for the groups which would result from a given partition. These calculations are computed for all possible partitions of each predictor variable. The algorithm then determines which potential partition of which predictor variable will result in the lowest total within-group variability (highest between-groups variability) of the criterion variable. The data set is then split into two subgroups based on the value of that single predictor variable. After each iteration, the within-group variability of each subgroup is examined, and the data subset with the largest variability is selected to be split at the next iteration. Again, the algorithm examines each possible partition of each predictor variable, including those used previously to split the data, to select the best. Thus, the best partition is selected for each iteration, since all possible partitions of every predictor variable have been studied (Refs 6:63; and 21:8-4). The same criteria defining "best" apply.

Stopping Criteria. Any of several stopping criteria may be used to halt the successive AID iterations. The analyst may specify the maximum number of terminal groups, and the algorithm ends when the specified number of data subgroups have been created. Alternatively, the minimum number of observations allowable in a terminal group may be stated. Here, no group would be split unless it contained at least twice the minimum number of observations, or if either of the resulting subgroups would have fewer than the minimum number of elements. When no groups can be split further, action terminates. In addition, the analyst may set a minimum value for the total sum-of-squares in a subgroup as a percentage of the

total sum-of-squares for all observations in the original data base. Further, the analyst can establish a minimum value by which the selected partition must reduce the total sum-of-squares. For these cases, if no possible partition of any group meets the condition, processing ceases (Ref 21:8-16-8-17).

Algorithm. Thus, for each iteration, AID

1. Selects the subset with the highest within-groups variability for splitting;

2. For this group, examines all predictor variables to find the partition which minimizes the within-group variability of the resulting subgroups;

3. Selects the best partition and splits the group accordingly, where the split would not violate any of the stopping criteria; and

4. Repeats until one or more of the stopping criteria apply to all terminal groups, at which time processing ceases.

When splitting the data according to the "best" partition identified in step 3 above would violate a stopping criterion, the next-best partition is selected. This repeats as required until a partition which splits the data within the stopping criteria is found, or until processing ceases because no such group is found (Ref 21:8-16).

AID Notation. (Ref 21:8-13-8-17). The theoretical AID model is represented as:

$$Y(mi) = \mu(i) + e(mi); \qquad m=1..n(i);\ i=1..g \tag{6}$$

where

Y(mi) is the mth criterion variable observation in the ith group

g is the number of existing groups

$\mu(i)$ is the group mean for the ith group

e(mi) is a random error term assumed to have normal distribution.

With each iteration, the data are partitioned into successively smaller subgroups. The value of g increases by one, and the observations in the groups just formed by the partition are renumbered to uniquely identify them with their groups. $\overline{Y(i)}$, the sample mean of the observations for the ith group is the estimated criterion mean; $\bar{Y}$ is the criterion sample mean for the total data base. Thus, the total variability for the population as a whole, at any processing stage is represented by TSS(T), the total sum-of-squares:

$$TSS(T) = \sum_{i=1}^{g} \sum_{m=1}^{n(i)} (Y(mi) - \bar{Y})^{**2} \qquad (7)$$

This value does not change across the iterations since it represents the sum-of-squares deviation from the overall mean of the total data base. The partitioned sum-of-squares, which follows algebraically from TSS, is written:

$$\sum_{i=1}^{g} \sum_{m=1}^{n(i)} (Y(mi) - \bar{Y})^{**2} = \sum_{i=1}^{g} \sum_{m=1}^{n(i)} (Y(mi) - \overline{Y(i)})^{**2} + \sum_{i=1}^{g} n(i)(\overline{Y(i)} - \bar{Y})^{**2} \qquad (8)$$

$$TSS(T) = WSS + BSS \qquad (9)$$

where

TSS(T) is the total sum-of-squares for the entire data base

WSS is the within-groups sum-of-squares

BSS is the between-groups sum-of-squares

Thus, the AID algorithm can be restated as minimizing WSS or maximizing BSS. Similarly, the total sum-of-squares for each group, TSS(i), is written:

$$TSS(i) = \sum_{i=1}^{g} (Y(mi) - \overline{Y(i)})^{**2} \qquad (10)$$

$$TSS(i) = BSS\ (i) + WSS(i) \qquad (11)$$

AID also produces an R-squared value analogous to the multiple R-squared value produced by multilinear regression and computed as:

$$R^{**2} = BSS(\text{all current groups})/TSS(T) \qquad (12)$$

An R-squared value close to one indicates that values of the criterion variable grouped by the algorithm are nearly identical. This suggests the predictor variables should accurately predict the value of the criterion variable.

AID Output. The AID analysis successively partitions a data set into mutually exclusive subgroups. One or more tree diagrams are normally produced to visually document the iteration action. The tree branches are defined by the predictor variables, and the leaves contain a variety of statistical data on the subgroups. Analyzing the tree contents and shape provides data on the interdependence of the criterion and predictor variables. Again, the subgroups derived by this algorithm best explain the variability

of the criterion variable with respect to the predictor variables (Refs 6:63+; and 21:8-51+).

CPESIM Data Processing

This thesis effort applied AID to the same data used in the multilinear regression procedure described in Chapter II. The dependent predictor and independent criterion variables specified in Table I (page 13) are also tested under AID. Again, both turnaround time and IO time were studied. All variables were considered to be ordinal, i.e., the values were treated in an ordered numerical sequence. This enabled the variable entries to be treated in groups, such as all jobs which required fewer than 54 memory blocks, rather than individually, such as all jobs which required exactly 23 memory blocks. The analyst could now better classify the requirements of certain types of jobs, for example, all jobs which required less than 54 memory blocks and only one tape drive.

Data Manipulation. Before AID was used, the data were manipulated into somewhat artificial groupings for each variable. AID checks each possible split of each predictor variable at each iteration for the selected (sub)group. If this check were performed between each value of each variable, rather than between groups, the processing time required would be massive. This processing is only complicated and increased still further by the number of variables AID can process in a single run. AID was designed to accept one criterion variable and up to and including 80 predictor variables for each run (Ref 6:308). Thus, to make the technique viable for large data files, the data were grouped before processing.

AID permits grouping the data in two methods. In the first, each predictor variable is defined to be composed of equal length intervals. The

data contained in each interval are considered and treated as a group. The second method also groups the data contained in the intervals, which here can be of differing lengths. The user specifies the length of each individual interval for the second method. In either case, no more than forty groups, or categories, are allowed (Refs 6:315-6; and 21:8-37-8-46).

In addition, AID accepts only integer data. Hence, data normally defined as fixed-point had to be converted by round-off, truncation, decimal point deletion, or value recoding (Ref 6:308). The last three methods were used in the successive AID runs made during this study.

The data were first grouped in variable-length intervals. Table IX contains the number of groups, and the size and range of each group for each variable. The number of groups ranged from four for tape-drives to 16 for several variables. Where feasible, the groups contained 40-60 elements. This figure was arbitrarily selected as a conveniently sized group to test, allowing for an adequate number of divisions while recognizing that the categories would contain many more data elements if a larger sized file was processed. The fractional parts of the predictor variables were truncated, except for turnaround time for which the decimal point was itself dropped. (The field was considered F11 rather than F11.4 and only then was recoded.) Another AID run was processed using fixed-length intervals. Here again the decimal values were truncated, except for turnaround time which was treated as specified above. Because the results of both groups were statistically similar, at the 95 percent significance level, and as tested by T-test, only that run involving variable length intervals was treated in detail.

AID Processing. AID was then run for turnaround time and IO time using

Table IX

Variable Length AID Intervals

Group	Cards		CPUtime	
	Range	Number	Range	Number
1	0 - 9	49	0 - 3	54
2	10 - 19	52	4 - 8	57
3	20 - 29	57	9 - 15	60
4	30 - 44	54	16 - 24	57
5	45 - 59	43	25 - 39	55
6	60 - 79	54	40 - 52	46
7	80 - 99	50	53 - 67	43
8	100 - 124	58	68 - 89	52
9	125 - 159	51	90 - 117	58
10	160 - 224	64	118 - 154	54
11	225 - 274	51	155 - 199	49
12	275 - 374	49	200 - 249	48
13	375 - 524	54	250 - 359	49
14	525 - 799	48	360 - 539	56
15	800 - last	58	540 - last	49
16	...	..	...	..

Table IX (continued)

Variable Length AID Intervals

Group	DiskIO Range	DiskIO Number	IOtime Range	IOtime Number
1	0 - 19	52	0.0000 - 1.9999	42
2	20 - 47	56	2.0000 - 5.4999	54
3	48 - 74	50	5.5000 - 9.9999	49
4	75 - 104	54	10.0000 - 14.9999	40
5	105 - 159	55	15.0000 - 29.9999	51
6	160 - 224	50	30.0000 - 74.9999	46
7	225 - 299	42	75.0000 - 114.9999	49
8	300 - 399	49	115.0000 - 149.9999	48
9	400 - 549	52	150.0000 - 199.9999	50
10	550 - 799	56	200.0000 - 249.9999	46
11	800 - 1099	55	250.0000 - 324.9999	53
12	1100 - 1499	55	325.0000 - 399.9999	41
13	1500 - 1999	51	400.0000 - 499.9999	60
14	2000 - 2999	57	500.0000 - 699.9999	55
15	3000 - last	53	700.0000 - 949.9999	51
16	...	..	950.0000 - last	52

Table IX (continued)

Variable Length AID Intervals

Group	Lines		Memory	
	Range	Number	Range	Number
1	0 - 11	47	0 - 12	29
2	12 - 29	53	13 - 16	47
3	30 - 49	51	17 - 19	51
4	50 - 74	43	20 - 27	57
5	75 - 124	43	28 - 30	46
6	125 - 174	44	31 - 34	54
7	175 - 299	48	35 - 53	45
8	300 - 449	46	54 - 59	58
9	450 - 599	49	60 - 63	46
10	600 - 849	54	64 - 65	46
11	850 - 1099	51	66 - 67	47
12	1100 - 1299	47	68 - 101	55
13	1300 - 1749	58	102 - 119	55
14	1750 - 2499	49	120 - 136	58
15	2500 - 3699	58	137 - 189	43
16	3700 - last	46	190 - last	50

Table IX (continued)

Variable Length AID Intervals

Group	TapeIO		Tapes	
	Range	Number	Range	Number
1	0	247	0	244
2	1 - 14	57	1	154
3	15 - 49	55	2	212
4	50 - 89	58	3	177
5	90 - 139	55	...	...
6	140 - 214	51	...	...
7	215 - 299	50	...	...
8	300 - 399	50	...	...
9	400 - 574	56	...	...
10	575 - 849	55	...	...
11	850 - last	53	...	...
12	...	...	...	...
13	...	...	...	...
14	...	...	...	...
15	...	...	...	...
16	...	...	...	...

those groups. Table X shows sample detailed output from the turnaround time run. As shown AID considered each possible split for the groups in the predictor variable. The group identifier, number of data elements, the sum of the Y(i), mean, standard deviation, BSS, and WSS of each potential resulting group were computed, considered, and printed. Also included were the maximum BSS and the BSS/TSS produced by the selected division, together with where that division would occur if selected. Similar details were produced for each variable. The analyst specifies the number of iterations which will produce detailed information. For this project, two iterations produced detailed data. When five or more were requested, the run required excessive IO time to print the results. Table XI contains a sample of the summary data produced for succeeding iterations. Summary data includes the variable actually chosen to be split, and where the split occurred. It also contains the total TSS, BSS, and WSS, the R-squared value, and other statistics for the selected division. Finally, information similar to that provided in the detailed output are included for the candidate groups remaining to be split. If the specific division selected fails any of the stopping criteria, this fact, together with the applicable stopping criterion, would also be shown, before the data on the division actually selected appears.

Two tree diagrams can also be requested through AID. Figure 4 contains a portion of the detailed tree produced for turnaround time. For each group produced, except the initial group, statistics include the group identifier, group mean, R-squared value, the number of elements in the group, the standard deviation, and the variable on which the division which produced this group occurred. The tree structures shown in Figure 5 for turnaround time and Figure 6 for IO time demonstrate the general shape of the final tree and

Table X

AID Sample Detailed Output

** STEP NO.= 1 PARENT GROUP = 1 **
TRY ON PREDICTOR 1 CPU-TIME-SECONDS

CODE	N	TOTAL WEIGHT	SUM OF Y	SUM Y-SQUARE	MEAN	STD. DEV.	B S S	W S S
0	54	54.000000	3112.0000	633026.80	57.629630	91.659858		
							486899.87	13610670.
1	57	57.000000	3461.0000	984997.80	60.719298	116.89388		
							872903.54	13144666.
2	60	60.000000	4862.0000	1123276.0	81.033333	110.24311		
							1213623.8	12803946.
3	57	57.000000	5295.0000	1302753.8	92.894737	119.27496		
							1485452.6	12532117.
4	55	55.000000	5685.0000	1485133.0	103.36364	127.74348		
							1709618.1	12307951.
5	46	46.000000	6009.0000	1551745.0	130.63043	129.10355		
							1710690.9	12306879.
6	43	43.000000	5060.0000	1271788.0	116.27907	126.71886		
							1875613.9	12142556.
7	52	52.000000	6280.0000	1287460.0	120.76923	100.86446		
							2097051.5	11920518.
8	58	58.000000	11189.000	3365983.0	192.91379	144.28503		
							1595506.7	12422063.
9	54	54.000000	11420.000	3350408.0	211.48148	131.60558		
							1062891.4	12954678.
10	43	43.000000	9060.0000	2365550.0	184.69796	118.69825		
							857699.45	13159870.
11	48	48.000000	8905.0000	2382325.0	185.52083	123.39486		
							687961.92	13329668.
12	49	49.000000	8897.0000	2481499.0	181.57143	132.94505		
							579671.69	13437898.
13	56	56.000000	10995.000	2962225.0	196.33929	112.88853		
							381383.94	13636186.
14	49	49.000000	11115.000	3300325.0	226.83673	126.13846		

MAX. BSS= 2097051.5 BSS/TSS = .14960 BETWEEN CODES 0 1 2 3 4 5 6 7
AND CODES 8 9 10 11 12 13 14

Table XI

AID Sample Summary Output

```
**********************************************************************************************************************
SPLIT GROUP   4 ON PREDICTOR   4   ASSIGNED-TAPES      INTO GROUP  18 WITH CODES  2
                                                      AND GROUP   19 WITH CODES  3

        BSS =      106791.28        BSS/TSS =     .07934   F-VALUE  2.94
**********************************************************************************************************************

                                                CURRENT SUMMARY


NCF      TOTAL TSS        TOTAL BSS        TOTAL WSS       R-SQUARED        R         F-RSQ     DF1   DF2      F-ANOVA     DF1   DF2
 18     14:17570.         7:90306.8        6527262.7        .53435133     .73899      12.7123     1    777       99.0711     9    777

**********************************************************************************************************************

**********************************************************************************************************************

GROUP 19 CANNOT BE SPLIT.THE NUMBER OF CASES IN ANY  SPLIT IS LESS THAN THE SPECIFIED MINIMUM

NUMBER OF CASES IN THIS GROUP=      46.
MINIMUM SPECIFIED BY USER    =      25  FOR ANY GIVEN SPLIT
**********************************************************************************************************************
```

GROUP	N	TOTAL WEIGHT	SUM OF Y	SUM Y-SQUARE	T S S	MEAN	STD. DEV.
19	46	46.000001	9437.0000	2767349.)	831327.93	205.15217	134.43343

CANDIDATE GROUPS ARE AS FOLLOWS.

GROUP	N	TOTAL WEIGHT	SUM OF Y	SUM Y-SQUARE	T S S	MEAN	STD. DEV.
12	85	85.000001	19835.000	5674575.0	1046319.4	233.35294	110.94882
15	60	60.000001	17020.000	5037450.0	979143.33	283.66667	127.76563
13	71	71.000001	14165.000	3658575.)	824571.83	199.78873	107.76682
16	126	126.00001	11977.000	1650383.0	521902.61	95.055556	64.359026
18	56	56.000000	7847.0000	1507443.0	407888.13	140.12500	85.344692
8	227	227.00001	4307.0000	305541.01	223821.84	18.973568	31.400626

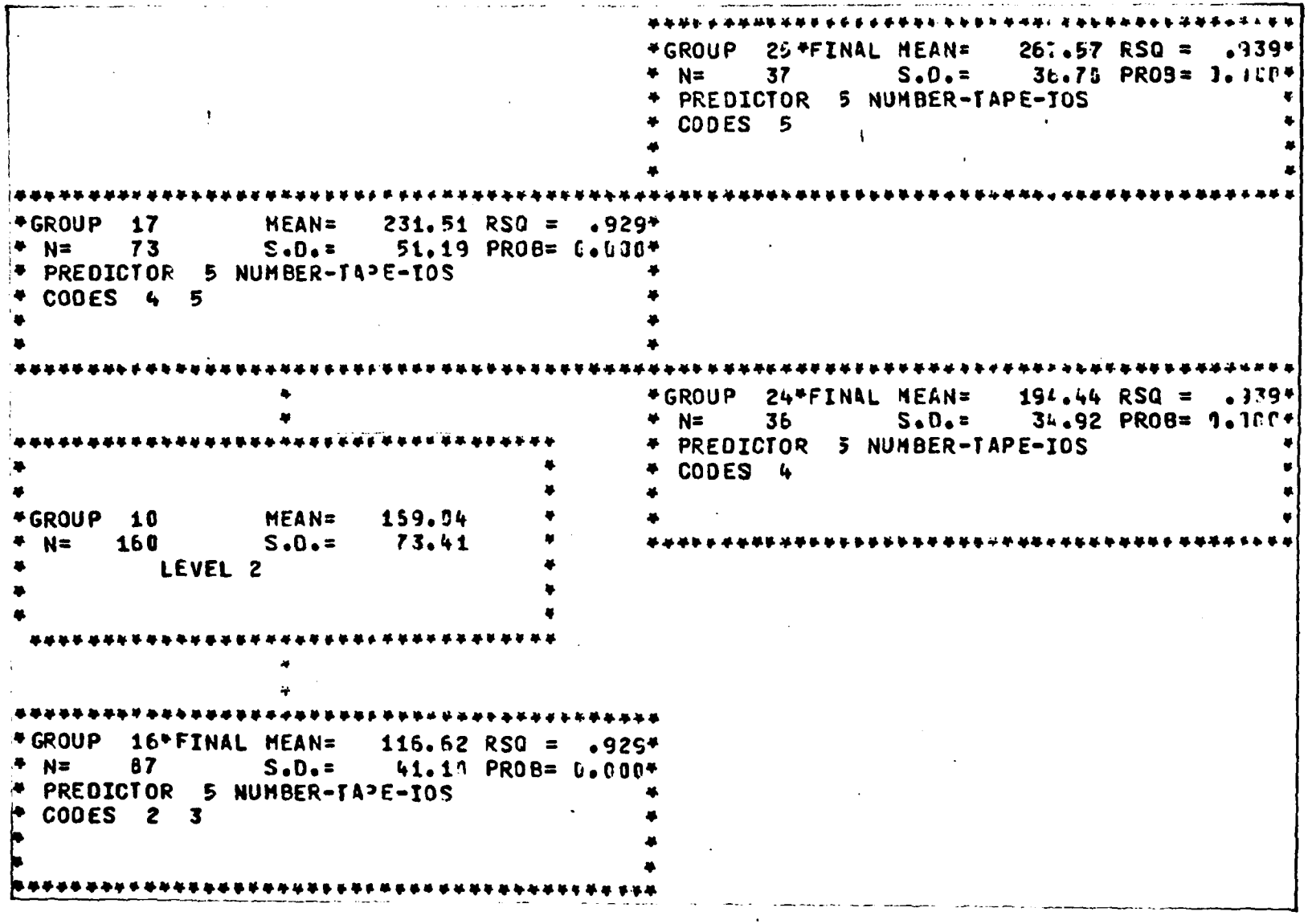


Figure 4. Sample AID Detailed Tree

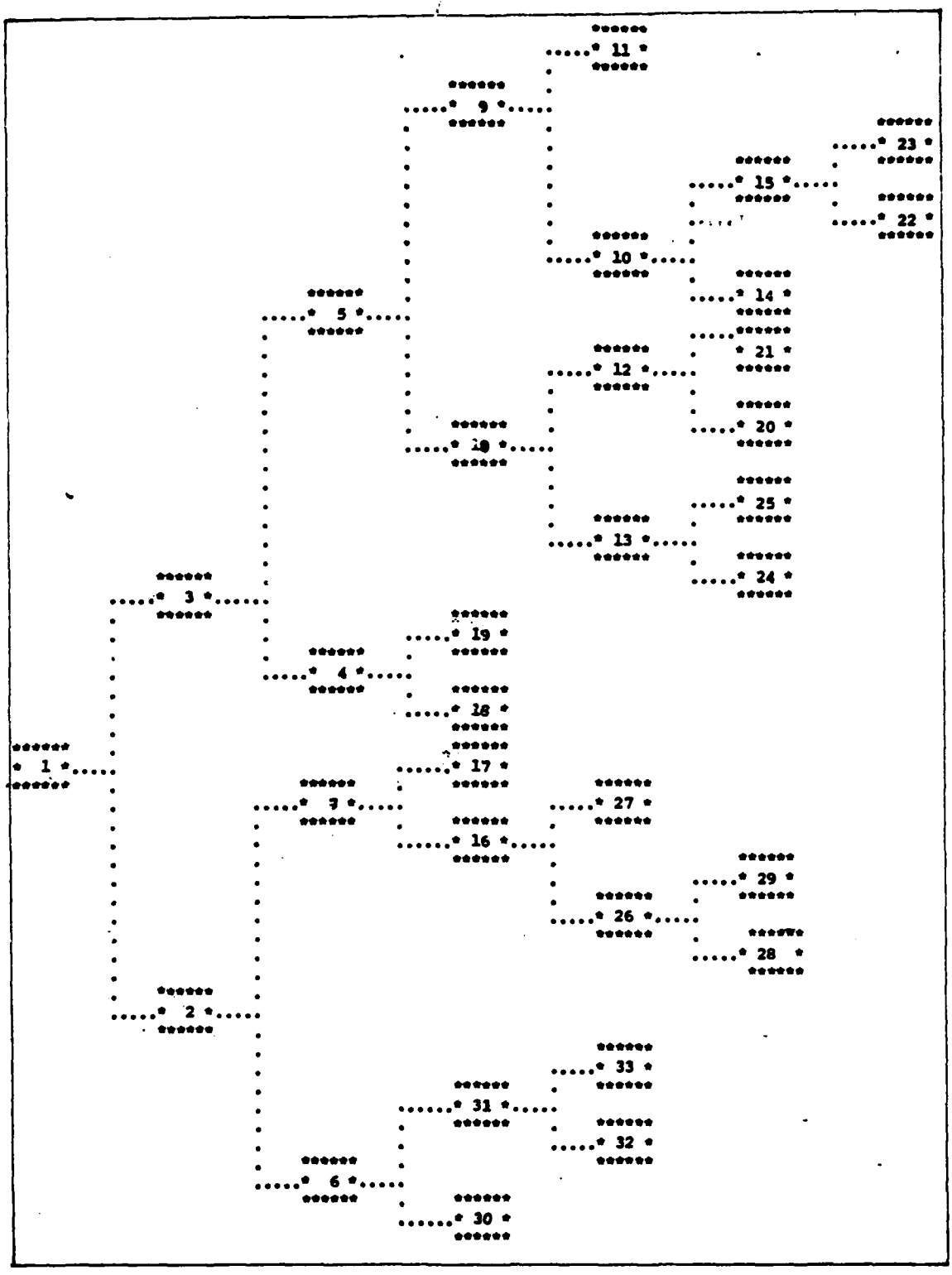


Figure 5. AID Turnaround Time Detailed Tree

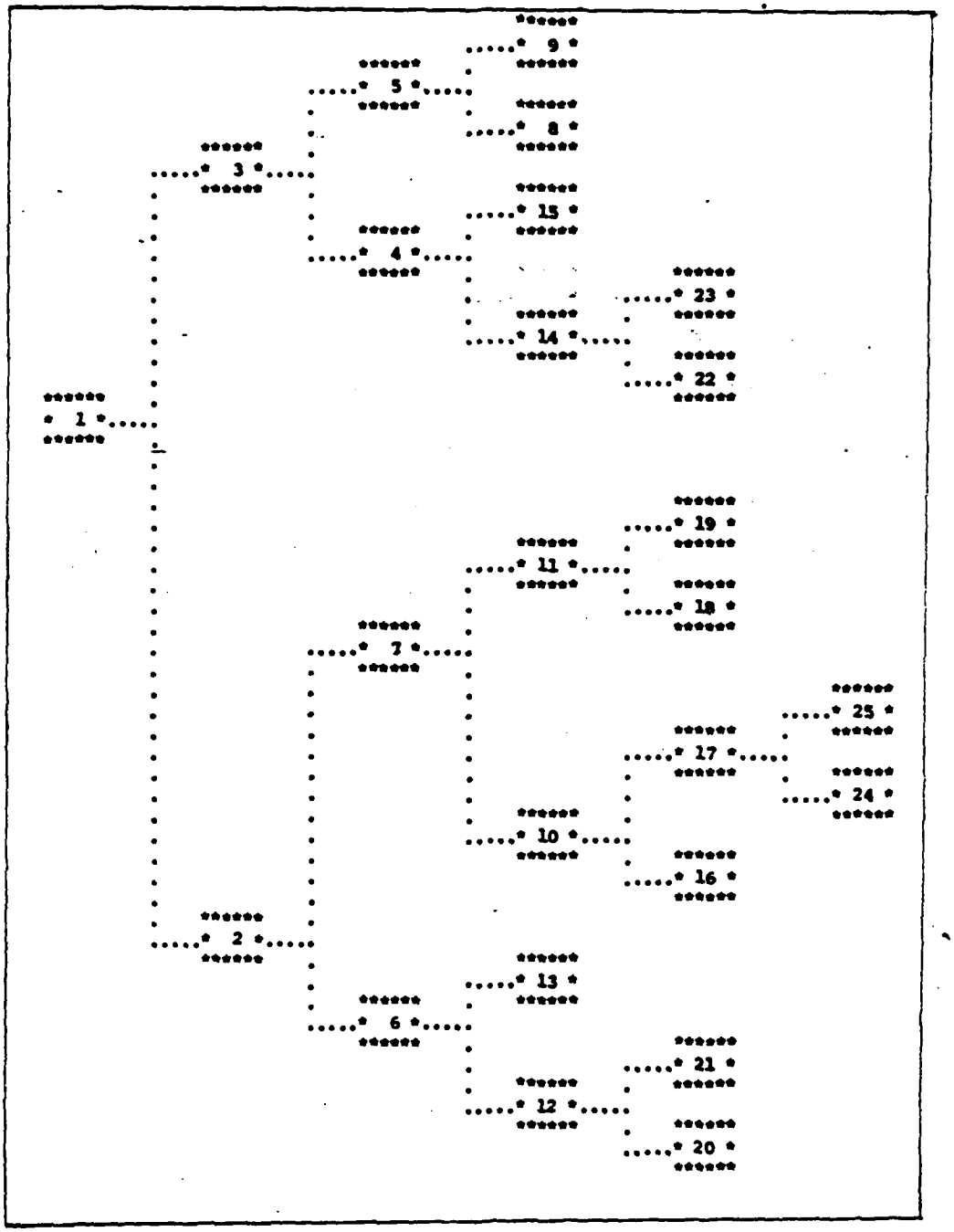


Figure 6. AID IO Time Detailed Tree

the order in which the branches were produced. For these two figures, the group identifier appears in each box. Further, the group with the higher mean of the pair produced at any division is shown as the top member of the pair.

Turnaround Time Model

Summary information on the successive divisions produced during turnaround time processing appears in Table XII. Data was first split on the variable tapes, between jobs which required two or three tape drives (Group 3) and those which needed fewer (Group 2). The first division of each of these groups was on the variable memory, between jobs which required 54 or more blocks of memory and those which needed fewer. After these divisions, the actions taken with the two major branches of the tree diverged significantly. Therefore, this level of the tree, called Level 3, was identified as a possible location to cut off the tree.

In addition, the R-squared value and its incremental change should be considered when selecting a level on which to trim the tree. Normally, splitting stops when successive divisions add little to the R-squared value (Ref 21:8-60-8-61). The final R-squared value of the upper branch (Group 3 descendents), if the tree had been trimmed after Level 3, was .410; its succeeding leaves produced R-squared values which ranged from .503-.543. Similarly, the final R-squared value of the lower branch (Group 2 descendents), if the tree had been trimmed at the specified point, was .474, as compared to leaf values which ranged from .527-.556. Although these figures reflected 25-32 percent and 11-17 percent increases, respectively, on the branches, Level 3 was still identified as a candidate for splitting.

The criterion variance for the subgroups produced must also be considered

Table XII

AID Summary Division Data
Turnaround Time Sequential Splits

Group Split	On Predictor	First Branch	Range	Second Branch	Range
1	Tapes	2	0 - 1	3	2 - 3
3	Memory	4	0 - 53	5	54 - End
2	Memory	6	0 - 53	7	54 - End
5	Tapes	8	2	9	3
9	Memory	10	54 - 136	11	137 - End
8	DiskIO	12	0 - 1099	13	1100 - End
10	Cards	14	0 - 79	15	80 - End
7	TapeIO	16	0 - 89	17	90 - End
4	Tapes	18	2	19	3
12	IOtime	20	0 - 249	21	250 - End
15	IOtime	22	0 - 324	23	325 - End
13	DiskIO	24	1100 - 2999	25	1000 - End
16	Tapes	26	0	27	1
26	DiskIO	28	0 - 799	29	800 - End
6	Tapes	30	0	31	1
31	DiskIO	32	0 - 159	33	160 - End

to decide on what level to trim the tree. Information produced by AID for each group includes the sample standard deviation for those specific data included in the subset. Preferably, this variance would be small for the final subgroups. When some final subgroup(s) contains a much larger variance than other subgroups, additional predictors may well be needed to fully explain the variance level of the group (Ref 21:8-61). As shown in Table XIII, Groups 5 and 4 of the upper branch have standard deviations of 124.61 and 114.87, respectively. The standard deviations of the leaves of this branch range from 96.68 to 132.15 for Group 5 descendents, and from 85.34 to 134.43 for Group 4 descendents. Similarly, Groups 7 and 6 of the lower branch have standard deviations of 88.78 and 31.40, respectively, while those of the leaves range from 53.18 to 119.64 for Group 7 descendents and from 15.60 to 39.98 for Group 6 descendents. Although the final variances differed significantly, from 15.60 to 134.43, those of the leaves of each of the four major subbranches thoroughly flanked the variance of the subbranch root, rather than being significantly less. Hence, while additional predictors seemed needed to fully explain the variance for all user jobs, except those requiring one or fewer tape drives and less than 54 memory blocks, those predictors were not available in the CPESIM data.

The initial AID divisions along tape drive and memory requirements mirrored exactly the variables included in the multilinear regression model. Further, stopping after these divisions did not significantly violate the other tree trimming criteria. Therefore, more detailed study was given to the four major subgroups produced by these early divisions.

A variety of statistical tests was performed on the four major subgroups. Tables XIV through XVII contain the basic statistical information

Table XIII

AID Detailed Tree Data
Turnaround Time

Group	Number of Elements	Mean	Standard Deviation	R-Square
1	787	141.400	133.460	.000
2	398	60.04	78.66	.380
3	389	224.66	126.49	.380
4	102	169.45	114.87	.410
5	287	244.28	124.81	.410
6	227	18.97	31.40	.474
7	171	114.54	88.78	.474
8	156	218.08	110.78	.491
9	131	275.47	132.72	.491
10	103	266.48	133.53	.503
11	28	345.36	102.97	.503
12	85	233.35	100.95	.506
13	71	199.79	107.77	.506
14	43	218.53	132.15	.514
15	60	283.67	127.77	.514
16	126	95.06	64.36	.527
17	45	169.11	119.64	.527
18	56	140.13	85.34	.534
19	46	215.15	124.43	.534

*Continued on next page

Table XIII (continued)

AID Detailed Tree Data
Turnaround Time

Group	Number of Elements	Mean	Standard Deviation	R-Square
20	43	219.53	116.63	.536
21	42	247.50	102.91	.536
22	27	252.41	122.84	.539
23	33	309.24	126.02	.539
24	44	176.93	96.68	.543
25	27	237.04	114.32	.543
26	72	78.71	62.92	.546
27	54	116.85	59.62	.546
28	45	64.36	53.18	.548
29	27	102.63	70.20	.548
30	172	8.11	15.60	.554
31	55	52.95	42.25	.554
32	30	35.77	35.94	.556
33	25	73.56	39.98	.556

Table XIV

Workload Parameter Statistics
Turnaround Time - Group 4

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	8.0 - 1738.0	334.716	334.791	268.957 - 400.475
CPUtime	0.0 - 999.0	207.863	206.671	167.269 - 248.457
DiskIO	44.0 - 7904.0	1168.569	1225.522	927.853 - 1409.284
IOtime	0.0 - 1800.0	519.697	488.684	431.567 - 607.627
Lines	1.0 - 8285.0	1441.725	1535.783	1140.069 - 1743.362
Memory	10.0 - 53.0	32.284	12.185	29.891 - 34.678
TapeIO	1.0 - 1673.0	437.863	449.293	349.613 - 526.112
Tapes	2.0 - 3.0	2.451	.500	2.353 - 2.549
Turntime	0.102 - 5.113	1.830	1.310	1.573 - 2.088

*Group 4 required two-three tape drives and less than 53 blocks of memory.

Table XV

Workload Parameter Statistics
Turnaround Time - Group 5

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 3146.0	309.986	405.172	262.911 - 357.061
CPUtime	0.0 - 999.0	13.174	223.184	197.610 - 249.471
DiskIO	4.0 - 9999.0	1201.362	1430.978	1035.105 - 1367.620
IOtime	0.0 - 3100.0	471.456	420.640	422.584 - 520.327
Lines	5.0 - 9090.0	1455.899	1483.296	1283.563 - 1628.235
Memory	54.0 - 205.0	100.425	44.820	95.218 - 105.633
TapeIO	2.0 - 2628.0	376.672	400.981	330.085 - 423.260
Tapes	2.0 - 3.0	2.456	0.499	2.398 - 2.514
Turntime	0.099 - 5.728	2.700	1.411	2.536 - 2.864

*Group 5 required two-three tape drives and more than 54 blocks of memory.

Table XVI

Workload Parameter Statistics
Turnaround Time - Group 6

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 1543.0	144.220	219.140	115.559 - 172.881
CPUtime	0.0 - 999.0	56.128	140.014	37.816 - 74.440
DiskIO	1.0 - 6769.0	332.370	776.149	230.859 - 433.881
IOtime	0.0 - 927.338	51.348	141.282	32.869 - 69.825
Lines	0.0 - 7156.0	278.141	890.959	161.614 - 394.668
Memory	10.0 - 53.0	23.687	10.715	22.286 - 25.089
TapeIO	0.0 - 832.0	25.744	109.095	11.476 - 40.013
Tapes	0.0 - 1.0	0.242	0.429	0.186 - 0.298
Turntime	0.006 - 1.549	0.212	0.311	0.171 - 0.252

*Group 6 required zero-one tape drives and fewer than 54 blocks of memory.

Table XVII

Workload Parameter Statistics
Turnaround Time - Group 7

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 1875.0	188.439	266.916	148.146 - 228.731
CPUtime	0.0 - 999.0	137.725	170.803	111.941 - 163.509
DiskIO	0.0 - 6343.0	966.456	1146.808	793.338 - 1139.575
IOtime	0.0 - 2000.0	221.358	289.247	177.694 - 265.021
Lines	2.0 - 6765.0	1148.351	1305.903	951.216 - 1345.486
Memory	54.0 - 205.0	99.550	47.443	92.388 - 106.712
TapeIO	0.0 - 1904.0	146.199	271.981	105.141 - 187.256
Tapes	0.0 - 1.0	0.579	0.495	0.504 - 0.654
Turntime	0.024 - 1.077	1.226	0.987	1.077 - 1.375

*Group 7 required zero-one tape drives and more than 53 blocks of memory.

for Groups 4 to 7, respectively. The cumulative information was produced by the SPSS Condescriptive procedure. Further, the initial file data in each group were compared to the corresponding data in the test file (defined in Chapter II for multilinear regression processing). The comparison was made using the SPSS T-Test procedure, and the results appear in Table XVIII. The only variable which was significantly different for any of the groups was turnaround time for Group 4, with two or three tape drives and less than 54 memory blocks required. For the initial data file, mean turnaround time was 1.8304; for the test file, it was 2.4110. None of the other variables for that group were significantly different, although two, tapes and tapeIOs, were close. Nevertheless, the mean turnaround time for this group fell within a range not overlapped by the other groups, as evidenced by the 95 percent confidence intervals.

Finally, the groups within a given file were compared to determine which variables, if any, were significantly different, other than memory and tape drives. Table XIX contains the results. The only groups which were not significantly different from each other were Groups 4 and 5 which required two or three tape drives. Here, the other predictor variables were statistically close for both memory groupings. However, these groups were statistically different from the groups requiring one or fewer tape drives.

Model. Thus the turnaround time model produced via AID would divide data into the four major groups documented in Tables XIV to XVII. Data meeting the group selection criteria would be expected to have turnaround time somewhere within the prediction interval 95 percent of the time, with the exception discussed above. However, data which fell in Group 4 would be expected to have turnaround time somewhere between that of Groups 5 and

Table XVIII

Turnaround Time T-Test Comparisons
Between Initial and Test Data Files

Variable	Group 4	Group 5	Group 6	Group 7
Cards	.349	.430	.566	.167
CPUtime	.176	.639	.287	.303
DiskIO	.217	.726	.153	.234
IOtime	.103	.639	.618	.261
Lines	.838	.631	.430	.412
Memory	.570	.539	.958	.570
Tape IO	.045	.994	.281	.418
Tapes	.035	.052	.752	.342
Turntime	.002*	.570	.036	.151

*The difference between the two files for the given group and variable was significant.

Table XIX

Turnaround Time T-Test Comparisons
Within Initial Data File

Variable	Groups 4 - 5	Groups 4 - 6	Groups 4 - 7	Groups 5 - 6	Groups 5 - 7	Groups 6 - 7
Card	.546	.000*	.000*	.000*	.000*	.079
CPUtime	.520	.000*	.000*	.000*	.000*	.000*
DiskIO	.825	.000*	.000*	.000*	.054	.000*
IOtime	.345	.000*	.000*	.000*	.000*	.000*
Lines	.936	.000*	.000*	.000*	.021*	.000*
Memory	.000*	.000*	.000*	.000*	.846	.000*
TapeIO	.226	.000*	.000*	.000*	.000*	.000*
Tapes	.925	.000*	.000*	.000*	.000*	.000*
Turntime	.000*	.000*	.000*	.000*	.000*	.000*

*The difference between the two groups considered was significant for the given variable.

7. Further, for all jobs which required zero or one tape drive, the values of the other prediction variables would be expected to fall within the prediction interval given for that variable and group 95 percent of the time. It should be noted the bounds of the prediction interval exceed those of the criterion interval for the given level of confidence.

IO Time Model

Summary information on the sequential divisions produced during IO time processing appear in Table XX. Data was first split on the variable tapeIO, between jobs which required less than 300 tapeIOs (Group 2), and those which required 300 or more (Group 3). Each of these groups was then divided further, still based on the variable tapeIO. Group 2 was split into Group 6 which needed only zero to 14 tapeIOs, and Group 7 which required 15 to 214. Group 3 was divided into Group 4 with 300 to 399 tapeIOs required, and Group 5 which needed 400 or more. The next divisions of each of these branches differed concerning the variable the split was based on. Therefore, this level of the tree, called Level 3, was identified as a possible location for trimming. One further aspect of the divisions was noted; all were based on either tapeIO or diskIO, the variables included in the regression equation.

The R-squared value and its incremental change were again considered. The final R-squared value of Group 3 descendents (the upper branch) was .788, if the tree had been trimmed after Level 3. Its succeeding leaves produced R-squared values which varied from .877 to .938. Similarly, the final R-squared value of Group 2 descendents (the lower branch) was .853, again if the tree had been trimmed after Level 3. The succeeding leaves on the lower branch had R-squared values which ranged from .896 to .939. Although these figures reflected 11-19 and 5-10 percent increases, respectively, on

Table XX

AID Summary Division Data
IO Time Sequential Splits

Group Split	On Predictor	First Branch	Range	Second Branch	Range
1	TapeIO	2	0 - 299	3	300 - End
3	TapeIO	4	300 - 399	5	400 - End
2	TapeIO	6	0 - 14	7	15 - 214
5	TapeIO	8	400 - 574	9	575 - End
7	DiskIO	10	0 - 1099	11	1100 - End
6	DiskIO	12	0 - 1499	13	1500 - End
4	DiskIO	14	0 - 549	15	550 - End
10	TapeIO	16	15 - 89	17	90 - 214
11	TapeIO	18	15 - 89	19	90 - 214
12	DiskIO	20	550 - 1499	21	0 - 549
14	TapeIO	22	215 - 299	23	300 - 399
17	TapeIO	24	90 - 139	25	140 - 214

the branches, Level 3 was still identified as a candidate for splitting.

The sample criterion variance and standard deviation for the subgroups was again also considered. As shown in Table XXI, Groups 4 and 5 (on the upper branch) had standard deviations of 98.54 and 172.44, respectively. The descendent leaves on this branch had standard deviations which ranged from 40.36 to 106.30 for Group 4 descendents, and from 95.10 to 156.50 for Group 5 descendents. Similarly, Groups 6 and 7 (the lower branch) had standard deviations of 69.72 and 111.75, respectively. The standard deviation of the leaves varied from 14.25 to 77.68 for Group 6 descendents, and from 41.10 to 267.57 for Group 7 descendents. Although these final variances also differed significantly, those of the leaves of three of the four major subbranches enclosed the variance of the subbranch root rather than being significantly less. The only exception was Group 5, although its variance was only 10.19 percent greater than that of its least explained leaf.

These initial AID divisions along tapeIO requirements directly reflected the more significant variable included in the multilinear regression model. Further, trimming the tree after Level 3 was produced did not significantly violate the other trimming criteria. Therefore, more detailed statistical tests were conducted on these four major subgroups.

Tables XXII through XXV contain the basic statistical information for Groups 4 to 7, respectively. Further, the initial file data was compared to the corresponding data in the test file (defined for multilinear regression in Chapter II). Again the comparison used the SPSS T-Test procedure. Results appear in Table XXVI. Two variables were significantly different, cards for Group 4 and tapes for Group 6. Neither variable was considered in the divisions leading to this model. The differences, therefore, should

Table XXI

AID Detailed Tree Data
IO Time

Group	Number of Elements	Mean	Standard Deviation	R-Square
1	787	243.02	273.79	.000
2	573	104.68	120.11	.684
3	214	613.44	220.43	.684
4	106	442.22	98.54	.788
5	108	781.49	172.44	.788
6	359	41.40	69.72	.853
7	214	210.85	111.75	.853
8	55	669.09	95.10	.877
9	53	898.13	156.50	.877
10	160	169.04	73.41	.896
11	54	334.72	114.00	.896
12	327	23.51	33.99	.916
13	32	224.22	77.68	.916
14	60	399.58	65.59	.920
15	46	497.83	106.30	.920
16	87	116.62	41.10	.929
17	73	231.51	51.19	.929
18	26	285.58	101.47	.931
19	28	380.36	105.72	.931

*Continued on next page

Table XXI (continued)

AID Detailed Tree Data
IO Time

Group	Number of Elements	Mean	Standard Deviation	R-Square
20	275	11.55	14.25	.935
21	52	86.75	37.89	.935
22	28	349.11	40.36	.938
23	32	443.75	49.61	.938
24	36	194.44	34.92	.939
25	37	267.57	36.70	.939

Table XXII

Workload Parameter Statistics
IO Time - Group 4

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	2.0 - 1384.0	262.123	297.588	204.811 - 319.435
DiskIO	5.0 - 9999.0	1036.226	1312.341	783.485 - 1288.968
IOtime	323.545 - 1100.0	492.775	112.856	471.040 - 514.510
Lines	5.0 - 7156.0	1445.896	1504.181	1156.109 - 1735.584
TapeIO	302.0 - 574.0	411.689	71.037	397.998 - 425.360
Tapes	1.0 - 3.0	2.142	0.710	2.005 - 2.278

*Group 4 required between 300 and 574 tapeIOs.

Table XXIII

Workload Parameter Statistics
IO Time - Group 5

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	9.0 - 3146.0	393.667	486.486	300.867 - 486.466
DiskIO	20.0 - 5957.0	1341.944	1401.705	1074.562 - 1609.327
IOtime	0.0 - 3100.0	1074.124	431.694	991.776 - 1156.471
Lines	13.0 - 9090.0	1417.648	1376.837	1155.010 - 1680.287
Tape IO	577.0 - 2628.0	981.269	402.042	904.577 - 1057.960
Tapes	1.0 - 3.0	2.222	0.702	2.088 - 2.356

*Group 5 required 575 or more tapeIOs.

Table XXIV

Workload Parameter Statistics
IO Time - Group 6

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 1529.0	153.153	211.993	131.150 - 262.126
DiskIO	0.0 - 6769.0	593.604	1007.916	488.989 - 698.220
IOtime	0.0 - 508.312	50.771	76.785	42.801 - 58.741
Lines	0.0 - 6765.0	533.900	1051.877	424.721 - 643.078
TapeIO	0.0 - 49.0	5.462	10.868	4.334 - 6.590
Tapes	0.0 - 3.0	0.501	0.849	0.413 - 0.589

*Group 6 required less than 50 tapeIOs.

Table XXV

Workload Parameter Statistics
IO Time - Group 7

Variable	Range	Mean	Standard Deviation	Confidence Interval
Cards	0.0 - 1996.0	293.288	371.920	243.273 - 343.503
DiskIO	4.0 - 7904.0	1106.650	1342.193	925.794 - 1287.505
IOtime	0.0 - 701.0074	239.993	118.786	223.987 - 255.999
Lines	2.0 - 8285.0	1525.065	1569.918	1313.525 - 1736.606
TapeIO	50.0 - 299.0	149.696	71.518	140.060 - 159.333
Tapes	1.0 - 3.0	2.159	0.765	2.056 - 2.262

*Group 7 required between 50 and 299 tapeIOs.

Table XXVI

IO Time T-Test Comparisons
Between Initial and Test Data Files

Variable	Group 4	Group 5	Group 6	Group 7
Cards	.023*	.865	.456	.643
DiskIO	.951	.747	.934	.645
IOtime	.500	.921	.773	.417
TapeIO	.251	.745	.065	.149
Tapes	.761	.258	.018*	.095

*The difference between the two files for the given group and variable was significant.

not have unduly biased the model. Further, the confidence intervals for Group 4 tapes were 0.413-0.589 and 0.563-0.744 for the initial and test files, respectively. The confidence intervals for Group 6 cards were 204.811-319.435 and 295.814-450.983 for the initial and test files, respectively. In both cases, the confidence intervals overlapped, and provided additional reason to consider the effect of these differences on the model minimal.

Finally, the groups within a given file were again compared to determine which variables, if any, were significant, discounting tapeIOs. The results appear in Table XXVII. Group 6, those jobs which involved less than 50 tapeIOs, was significantly different from the other groups in virtually every variable. In addition, Group 5, those jobs which required 575 or more tapeIOs, and Group 7, those jobs which needed 50 to 299 tapeIOs, were also significantly different.

Model. The IO time model produced using the AID algorithm divided data into the four major groups documented in Tables XXII to XXV. Data meeting the group selection criteria would be expected to complete processing with an IO time falling somewhere within the confidence interval 95 percent of the time. Further, for all jobs processing fewer than 50 tapeIOs, the values of the other prediction variables would be expected to fall within the confidence interval given for that variable and group 95 percent of the time.

Table XXVII

IO Time T-Test Comparisons
Within Initial Data File

Variable	Groups 4 - 5	Groups 4 - 6	Groups 4 - 7	Groups 5 - 6	Groups 5 - 7	Groups 6 - 7
Cards	.018*	.467	.417	.000*	.000*	.000*
DiskIO	.101	.000*	.654	.000*	.627	.000*
IOtime	.000*	.000*	.000*	.000*	.000*	.000*
Lines	.886	.000*	.663	.000*	.000*	.827
TapeIO	.000*	.000*	.000*	.000*	.000*	.000*
Tapes	.404	.000*	.841	.000*	.487	.000*

*The difference between the two groups considered was significant for the given variable.

IV. Ridge Regression

Theoretical Discussion

Ridge regression is an alternative estimation method which proves most useful in explaining the interrelationships between a criterion variable and several highly nonorthogonal, or correlated, predictor variables. Nonorthogonal predictor variables may be highly multicollinear, and thus may produce estimation errors in multilinear regression by pulling the least squares estimates of the regression coefficients away from the true values being estimated. The coefficients derived through multilinear regression may therefore be too large in magnitude, or of the wrong sign. Furthermore, the least squares solution used by multilinear regression is unstable. Slight variations in the data can result in significantly different estimates of the coefficients (Refs 3:181; 13:82; 14:55; and 28:263). Ridge regression often overcomes or circumvents these problems.

This technique uses a diagnostic tool called a ridge trace as an orderly approach to show the complex interrelationships existing between nonorthogonal predictor variables, and the effects of these interrelationships on estimation. This two-dimensional plot provides a readily interpreted picture of the interrelationships. The estimators produced by ridge regression are admittedly biased, but the method accepts a small amount of bias to achieve a major variance reduction. Further, this technique helps circumvent the sensitivity of the estimates to a particular data set, since slight variations in the data do not affect ridge estimates as severely as least squares estimates. Thus, the estimates produced by ridge regression are also closer to the accurate regression coefficients. Ridge regression is helpful specifically when multicollinearity of data is suspected, other-

wise it will produce less accurate results. This technique presents an alternative data interpretation to perhaps better explain the data base being studied. The method is not a panacea, and may sometimes produce less accurate results; nevertheless, it frequently proves most valuable (Refs 3:181; 13:70+; 14:55; and 20:591).

Mathematical Development. Ridge estimates may be defined and computed using various methods; Chatterjee discusses several in detail (Ref 3:189-192). Ridge regression also produces a linear equation relating the value of the criterion variable to those of the predictor variables. The specific ridge algorithm used in this thesis effort computes a set of solutions defined as (Ref 26:1):

$$\underline{b^*} = ((X'X + kI)^{**}(-1))X'\underline{y} \qquad (13)$$

where

$\underline{b^*}$ is a column vector of dimension m containing the set of standardized (normalized) regression coefficient estimates

X'X is the correlation matrix for the m independent variables

k is the bias factor

X'$\underline{y}$ is the correlation vector of the dependent and independent variables

$\underline{y}$ is a column vector of dimension n containing the dependent variable observations.

The bias factor, k, is iteratively varied to create regression coefficient estimates with differing amounts of bias. When k = 0, the estimators are the ordinary least squares estimates. As k increases, ($0 <= k <= 1$), the bias of the estimate increases. However, an optimum value of k normally exists for which the ridge estimates will stabilize and the system will

behave as if it were orthogonal. For this optimum k, the regression coefficients for those factors representing rates of change will have reasonable magnitudes, and all coefficients will have their proper signs. Further, the residual sum-of-squares will not be inflated to an unreasonable value relative to either the minimum residual sum-of-squares or the reasonable variance involved in generating the data (Refs 3:181-2; and 14:65). Normally the optimum value of k is selected after computing $\underline{b}^*$ for a range of k values. The results are plotted on a ridge trace, and the graph, normalized and unnormalized regression coefficients, and variance inflation factors are analyzed to select the appropriate k. Estimates based on $\underline{b}^*$ are biased, thus implying a prior bound on the regression vector. The ridge trace presents the complex interrelationships existing between the nonorthogonal prediction variables and the effect of these interrelationships on the estimators of $\underline{b}^*$ (Refs 3:182; 13:70; and 14:66). In addition, although ridge regression is not specifically designed to select variables, those variables whose coefficients rapidly approach zero as k increases tend to be inherently deleted from consideration. A matter of interest, therefore, is now small the $\underline{b}^*$ must be to delete a given term (Ref 12:11+).

Detection of Multicollinearity. Two methods are commonly used to detect multicollinearity for ridge regression. The first method considers how multicollinearity affects the error between the ordinary least squares estimators and the true regression coefficients determined by ridge regression. This method calculates a variance inflation factor (VIF) associated with each coefficient which represents the amount to which multicollinearity inflates the variance of the coefficient. For each $\underline{b}^*$ (Ref 3:182):

$$VIF = (1 - R^{**}2)^{**}(-1) \qquad (14)$$

where

R**2 is the square of the multiple correlation coefficient of the ith predictor variable produced by regressing that variable against all other predictor variables.

As R**2 increases toward one, the VIF for the estimated coefficient of a given variable tends to infinity, and multicollinearity becomes highly suspect. A VIF greater than ten suggests multicollinearity may be the source of estimation problems (Refs 3: 182-3; and 12:28-31).

The second method considers the basic instability of ordinary least squares estimators to slight changes in the data. This instability, when it exists, may be demonstrated by plotting the regression coefficients against k, the bias factor. When small values of k, k=0.001, etc., are used, the ridge estimates obtained may be considered as the results of slightly altered data. Large fluctuations in the estimates for small values of k demonstrate instability which may well be caused by multicollinearity (Refs 3:183; and 12:28-31).

Algorithm. Therefore, the analyst normally creates a ridge trace by processing the data with varying values of k, and plotting the interaction of the independent variables against the corresponding k values. This trace is used to select a value of k and the regression coefficient estimates which correspond. The k selected will stabilize the system and make it approximate orthogonal behavior. In addition, the other guidelines discussed above will be met. The smallest suitable k should be selected to minimize the bias introduced (Refs 3:185; and 14:65).

Ridge Output. The ridge regression algorithm used in this thesis produces the sample means and standard deviations of all variables considered, together with the sample covariance matrix for all variables. It also produces values for both the standardized and unnormalized coefficient estimates. Further, it creates the ridge trace, plotting the standardized coefficient values determined against the values of k. Finally, it calculates the values of the VIFs for each coefficient at each k value treated. These data, particularly the ridge coefficients, can then be interpreted to better understand the system under study.

CPESIM Data Processing

This thesis effort applied ridge regression to the same data used in the multilinear regression procedure described in Chapter II. The dependent predictor and independent criterion variables specified in Table I (page 13) were also tested under ridge regression. Again, both turnaround time and IO time were studied.

Ridge Regression Processing. Ridge regression was then run for turnaround time and IO time. Several (12-16) ridge runs were processed for each criterion variable. Examples of typical output appear in the discussion of each model. The values of k ranged from 0.0020 to 0.0100 for each. (For the ridge program used, k can vary only from 0.0001 to 0.0200.) For each variable, the approximate vicinity in which k stabilized was identified in the first runs. Subsequent processing obtained a more precise, optimal value for k.

Turnaround Time Model

The correlation matrix for the variables used appears in Table XXVIII. Again, turnaround time was most highly correlated with the variables tapes

Table XXVIII

Turnaround Time Correlation Matrix

Variable	CPUTime	Memory	IOTime	Tapes	Turntime	Cards	Lines	DiskIO	TapeIO
CPUTime	1.000	.182	.297	.341	.291	.228	.261	.159	.267
Memory	.182	1.000	.231	.311	.463	.029	.233	.131	.207
IOTime	.297	.231	1.000	.457	.369	.255	.272	.382	.969
Tapes	.341	.311	.457	1.000	.660	.203	.301	.192	.440
Turntime	.291	.463	.369	.560	1.000	.119	.203	.171	.344
Cards	.228	.029	.255	.263	.119	1.000	.144	.221	.213
Lines	.261	.233	.272	.301	.208	.144	1.000	.185	.225
DiskIO	.159	.131	.382	.192	.171	.221	.185	1.000	.174
TapeIO	.267	.207	.969	.440	.344	.213	.225	.174	1.000

and memory, in that order. However, multicollinearity is not a factor since turnaround time is a dependent variable. However, multicollinearity again seemed to exist between IOtime and tapeIO, since the correlation coefficient was 0.969. Since these variables were not included in the model, the correlation would not bias the model.

Tables XXIX and XXX contain the normalized (standardized) and unnormalized regression coefficients, respectively. The normalized coefficients for variables cards and lines never changed sign, but these variables were excluded from the model since the coefficients were rapidly approaching zero as k increased. (These variables were therefore documented in no further tables for this model.) The coefficient for tapeIO changed sign near k = 0.081. At this point, the other coefficients had also begun to stabilize, making this the earliest k for which the system could be considered to be approximately orthogonal. However, by this point, as for virtually the entire range, only variables tapes and memory could affect the regression equation, since the other unnormalized coefficients remained equal to zero. This condition also indicated the minimal effect the other variables had on turnaround time.

The ridge trace for the normalized coefficients appears in Figure 7 while Table XXXI contains the VIFs for the regression coefficients. The VIFs for IOtime and tapeIO at k = 0 were very high, 76.7 and 66.7, respectively. These values also indicated possible multicollinearity between these variables. Again, however, neither were included in the regression equation, so this condition did not unduly bias the model. The ridge trace seemed to stabilize for k approximately 0.084. However, only at k = 0.108 had all the VIFs reached unity, a condition characteristic of an orthogonal system. For this value also, the improper negative sign for tapeIO had

Table XXIX

Turnaround Time Normalized Regression Coefficients

K-Value	CPUTime	Memory	IOTime	Tapes	Cards	Lines	DiskIO	TapeIO
0.000	.030	.279	.453	.551	-.023	-.057	-.071	-.378
.009	.053	.279	.210	.543	-.021	-.050	-.018	-.109
.018	.054	.278	.144	.537	-.020	-.047	-.004	-.086
.027	.055	.276	.113	.531	-.020	-.045	-.003	-.055
.036	.056	.275	.096	.525	-.019	-.043	.007	-.037
.045	.057	.273	.085	.519	-.019	-.041	.010	-.025
.054	.057	.272	.077	.514	-.018	-.039	.011	-.017
.063	.058	.270	.072	.508	-.018	-.037	.013	-.010
.072	.058	.269	.068	.503	-.017	-.035	.014	-.005
.081	.059	.267	.065	.498	-.017	-.034	.015	-.000
.090	.059	.266	.063	.493	-.016	-.033	.016	.003
.099	.060	.264	.061	.488	-.016	-.031	.016	.007
.102	.060	.264	.061	.485	-.015	-.031	.016	.008
.105	.060	.263	.060	.485	-.015	-.031	.017	.009
.108	.060	.263	.060	.483	-.015	-.030	.017	.009
.111	.060	.262	.059	.482	-.015	-.029	.017	.010
.114	.061	.262	.059	.480	-.015	-.029	.017	.011
.117	.061	.261	.059	.478	-.015	-.029	.017	.012
.126	.061	.260	.058	.474	-.014	-.027	.018	.014
.135	.061	.258	.057	.469	-.014	-.026	.018	.016
.144	.062	.257	.057	.465	-.013	-.025	.018	.018

Table XXX

Turnaround Time Unnormalized Regression Coefficients

K-Value	Intercept	CPUTime	Memory	IOTime	Tapes	DiskIO	TapeIO
0.000	-.7920E-01	.000	.008	.002	.716	-.000	-.002
.009	-.7128E-01	.000	.008	.001	.706	-.000	-.001
.018	-.6277E-01	.000	.008	.001	.698	-.000	-.000
.027	-.5424E-01	.000	.008	.000	.690	.000	-.000
.036	-.4579E-01	.000	.008	.000	.682	.000	-.000
.045	-.3746E-01	.000	.008	.000	.674	.000	-.000
.054	-.2923E-01	.000	.008	.000	.667	.000	-.000
.063	-.2112E-01	.000	.008	.000	.660	.000	-.000
.072	-.1312E-01	.000	.008	.000	.653	.000	-.000
.081	-.5238E-02	.000	.008	.000	.647	.000	-.000
.090	.2533E-02	.000	.008	.000	.640	.000	.000
.099	.1020E-01	.000	.008	.000	.634	.000	.000
.102	.1273E-01	.000	.008	.000	.632	.000	.000
.105	.1525E-01	.000	.008	.000	.630	.000	.000
.108	.1775E-01	.000	.008	.000	.628	.000	.000
.111	.2025E-01	.000	.008	.000	.626	.000	.000
.114	.2273E-01	.000	.008	.000	.624	.000	.000
.117	.2521E-01	.000	.008	.000	.622	.000	.000
.126	.3256E-01	.000	.008	.000	.616	.000	.000
.135	.3981E-01	.000	.008	.000	.610	.000	.000
.144	.4697E-01	.000	.007	.000	.604	.000	.000

K-VALUE	.....	R-SQUARE
0.000	9 87 6. 1 2 3 4	.5177
0.003	9 786. 1 2 3 4	.5175
0.006	9 786. 1 3 2 4	.5172
0.009	9 7 8. 1 3 2 4	.5169
0.012	9 7 8. 1 3 2 4	.5167
0.015	9 7 8. 1 3 2 4	.5165
0.018	9 7 68 1 3 2 4	.5163
0.021	97 68 1 3 2 4	.5161
0.024	97 68 1 3 2 4	.5160
0.027	9 68 1 3 2 4	.5159
0.030	9 68 1 3 2 4	.5157
0.033	968 1 3 2 4	.5156
0.036	968 1 3 2 4	.5155
0.039	968 1 3 2 4	.5154
0.042	968 1 3 2 4	.5152
0.045	968 13 2 4	.5151
0.048	798 13 2 4	.5150
0.051	798 13 2 4	.5148
0.054	798 13 2 4	.5147
0.063	79.8 13 2 4	.5142
0.072	7698 13 2 4	.5138
0.081	7698 3 2 4	.5133
0.090	7698 3 2 4	.5128
0.099	7698 3 2 4	.5122
0.102	7698 3 2 4	.5121
0.105	7698 3 2 4	.5119
0.108	7698 3 2 4	.5117
0.111	7698 3 2 4	.5115
0.114	7698 3 2 4	.5113
0.117	7698 3 2 4	.5111
0.126	76.9 3 2 4	.5104
0.135	76.9 3 2 4	.5098
0.144	7.9 3 2 4	.5091

Figure 7. Turnaround Time Ridge Trace

Table XXXI

Turnaround Time Variance Inflation Factors

K-Value	CPUTime	Memory	IOTime	Tapes	DiskIO	TapeIO
0.000	1.2	1.2	76.7	1.4	4.6	66.7
.009	1.2	1.1	10.5	1.4	1.7	12.8
.018	1.2	1.1	6.1	1.4	1.3	5.4
.027	1.1	1.1	3.4	1.3	1.2	3.1
.036	1.1	1.1	2.2	1.3	1.1	2.1
.045	1.1	1.0	1.6	1.3	1.0	1.5
.054	1.1	1.0	1.3	1.2	1.0	1.2
.063	1.0	1.0	1.0	1.2	1.0	1.0
.072	1.0	1.0	.9	1.2	.9	.9
.081	1.0	1.0	.7	1.1	.9	.8
.090	1.0	.9	.7	1.1	.9	.7
.099	1.0	.9	.6	1.1	.9	.6
.102	.9	.9	.6	1.1	.9	.6
.105	.9	.9	.6	1.1	.9	.6
.108	.9	.9	.5	1.0	.9	.6
.111	.9	.9	.5	1.0	.9	.6
.114	.9	.9	.5	1.0	.9	.5
.117	.9	.9	.5	1.0	.8	.5
.126	.9	.9	.5	1.0	.8	.5
.135	.9	.8	.4	1.0	.8	.5
.144	.9	.8	.4	1.0	.8	.5

disappeared, and all included normalized coefficients had stabilized. Further, the R-squared value had decreased only 0.0054, from 0.5171 to 0.5117. Therefore, the solution at $k = 0.108$ was identified for the regression equation.

Model. The model for this solution was then:

$$\text{Turnaround Time} = 0.01775 + 0.628*\text{Tapes} + 0.008*\text{Memory} \qquad (15)$$

However, since the R-squared value was only 0.5171, the model was not expected to explain the variation in turnaround time to a truly significant degree. Therefore, the data records selected in Chapter II from the test data file to verify the multilinear regression model were again used to document the accuracy of this model. Table XXXII contains the predictor values used, the actual and calculated values of the criterion variable, and the percentage error between them.

The percent error ranged from -51.23 to 1984.05. The average relative error was 283.06 percent, with an average absolute error equaling 307.03 percent. The largest two values again seemed to be outliers, falling fully 844.75-1495.33 percentage points (172.85-305.97 percent) beyond the next greatest value. In addition, the actual values were 0.0124 with a calculated value of 0.17775, and 0.0116 with an estimated value of 0.24175. The relative increases were therefore 0.17651 and 0.24059, respectively, though the percentage increases were high. However, even with these two errors omitted, the mean relative error was 53.77 percent and the mean absolute error was 81.74 percent. This confirmed that equation 15 was inadequate to model turnaround time, and further effort was required to fully explain the variation in turnaround time. Factors measured in the simulated accounting log were inadequate to provide the necessary explanation.

Table XXXII

Turnaround Time Ridge Regression Evaluation Statistics

Memory	Tapes	Turnaround Time		Percent Error
		Actual	Calculated	
14	2	2.3121	1.38575	40.07
20	0	0.0124	0.17775	1333.47
64	3	1.5448	2.41375	56.25
116	2	1.0676	2.20175	106.23
54	0	0.1243	0.44975	161.83
28	1	0.9913	0.86975	- 12.26
27	0	0.0429	0.23375	448.72
14	2	2.1271	1.38575	- 34.85
51	2	3.4483	1.68175	- 51.23
149	3	4.3499	3.09375	- 28.88
28	0	0.0116	0.24175	1984.05
32	2	1.5332	1.52975	- 1.22
59	3	2.4613	2.37375	- 3.56
50	2	2.6472	1.67375	- 36.77

IO Time Model

The correlation matrix for the variables used to model IO time appears in Table XXXIII. Again, IO time was extremely highly correlated with tapeIO, with a correlation coefficient of 0.969. Multicollinearity was not indicated, however, since IO time was here used as the criterion variable.

Tables XXXIV and XXXV contain the normalized (standardized) and unnormalized regression coefficients, respectively. No coefficient had an improper sign, and all were fairly stable across the entire range of values of k. Thus these coefficients did not indicate a specific potential value of k for the model solution. Further, all the unnormalized coefficients were nonzero and would be included in the equation.

The ridge trace for the normalized coefficients produced appears in Figure 8. Table XXXVI contains the VIFs for the regression coefficients. The ridge trave seemed to stabilize for values of k close to 0.077. However, all the VIFs did not reach unity until $k = 0.082$. The coefficients for this value had stabilized. Further, the R-squared value had decreased by only 0.0060 from 0.9869 to 0.9808. Therefore, the solution at $k = 0.082$ was identified for the IO time regression equation.

Model. The model for this solution could then be written:

$$\text{IO time} = 5.673 + 11.210*\text{Tapes} + 0.022*\text{Cards} + 0.008*\text{Lines} + 0.064*\text{DiskIO} + 0.899*\text{TapeIO} \qquad (16)$$

Since the R-squared value was high, 0.9808, the model was expected to explain most of the variation in IO time. Again, the data records used in Chapter II to verify multilinear regression were used to determine the accuracy of this ridge regression model. Table XXXVII contains the results.

Table XXXIII

IO Time Correlation Matrix

Variable	IOTime	Tapes	Cards	Lines	DiskIO	TapeIO
IOTime	1.000	.457	.255	.272	.382	.969
Tapes	.457	1.000	.203	.301	.192	.440
Cards	.255	.203	1.000	.144	.221	.213
Lines	.272	.301	.144	1.000	.185	.225
DiskIO	.382	.192	.221	.185	1.000	.174
TapeIO	.969	.440	.213	.225	.174	1.000

AD-A115 850 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 9/2
APPLICATION OF STATISTICAL ANALYSIS TECHNIQUES TO COMPUTER PERF--ETC(U)
DEC 81 M A STOVER
UNCLASSIFIED AFIT/GCS/EE/81D-17 NL

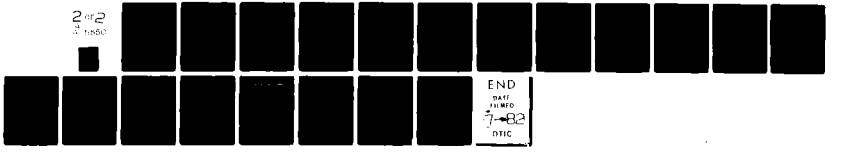


Table XXXIV

IOTime Normalized Regression Coefficients

K-Value	Tapes	Cards	Lines	DiskIO	TapeIO
0.000	.001	.007	.023	.215	.925
.008	.004	.008	.023	.214	.916
.015	.007	.009	.024	.213	.908
.023	.011	.010	.025	.212	.900
.030	.014	.011	.026	.211	.892
.038	.017	.013	.026	.209	.884
.045	.020	.014	.027	.208	.876
.053	.023	.015	.028	.207	.868
.060	.025	.016	.028	.206	.861
.067	.028	.017	.029	.205	.854
.075	.031	.018	.030	.204	.846
.077	.031	.018	.030	.204	.844
.081	.032	.018	.030	.203	.842
.082	.033	.019	.030	.203	.839
.085	.034	.019	.030	.203	.837
.087	.035	.019	.031	.202	.835
.090	.036	.019	.031	.202	.832
.097	.038	.020	.031	.201	.826
.105	.040	.021	.032	.200	.819
.113	.042	.022	.033	.199	.813
.120	.044	.023	.033	.198	.805

Table XXXV

IOTime Unnormalized Regression Coefficients

K-Value	Intercept	Tapes	Cards	Lines	DiskIO	TapeIO
0.000	1.879	.171	.008	.006	.067	.991
.008	2.132	1.363	.009	.006	.067	.982
.015	2.406	2.513	.011	.007	.067	.973
.023	2.701	3.822	.012	.007	.066	.964
.030	3.015	4.692	.013	.007	.066	.955
.038	3.348	5.724	.015	.007	.066	.947
.045	3.697	6.720	.016	.007	.065	.939
.053	4.063	7.682	.017	.008	.065	.930
.060	4.445	8.610	.018	.008	.065	.922
.067	4.841	9.507	.019	.008	.064	.915
.075	5.250	10.373	.021	.008	.064	.907
.077	5.390	10.655	.021	.008	.064	.904
.080	5.531	10.934	.021	.008	.064	.902
.082	5.673	11.210	.022	.008	.064	.899
.085	5.817	11.483	.022	.008	.063	.897
.087	5.962	11.752	.022	.008	.063	.894
.090	6.109	12.819	.023	.009	.063	.892
.097	6.556	12.801	.024	.009	.063	.885
.105	7.014	13.557	.025	.009	.063	.878
.113	7.482	14.288	.026	.009	.062	.871
.120	7.961	14.995	.027	.009	.062	.864

K-VALUE	.....			R-SQUARE
0.000	34	5	6	.9868
0.008	34	5	6	.9868
0.015	34	5	6	.9866
0.023	34	5	6	.9863
0.030	34	5	6	.9859
0.038	34	5	6	.9854
0.045	34	5	6	.9849
0.053	34	5	6	.9842
0.060	34	5	6	.9835
0.067	34	5	6	.9826
0.075	34	5	6	.9818
0.077	34	5	6	.9815
0.080	34	5	6	.9811
0.082	34	5	6	.9808
0.085	34	5	6	.9805
0.087	4	5	6	.9802
0.090	4	5	6	.9798
0.097	42	5	6	.9787
0.015	42	5	6	.9773
0.113	42	5	6	.9765
0.120	42	5	6	.9752

Figure 8. IO Time Ridge Trace

Table XXXVI

IOTime Variance Inflation Factors

K-Value	Tapes	Cards	Lines	DiskIO	TapeIO
0.000	1.3	1.1	1.1	1.1	1.3
.008	1.3	1.1	1.1	1.1	1.3
.015	1.3	1.1	1.1	1.1	1.2
.023	1.2	1.0	1.1	1.0	1.2
.030	1.2	1.0	1.1	1.0	1.2
.038	1.2	1.0	1.0	1.0	1.2
.045	1.2	1.0	1.0	1.0	1.1
.053	1.1	1.0	1.0	1.0	1.1
.060	1.1	1.0	1.0	1.0	1.1
.067	1.1	.9	1.0	.9	1.1
.075	1.1	.9	1.0	.9	1.0
.077	1.1	.9	.9	.9	1.0
.080	1.0	.9	.9	.9	1.0
.082	1.0	.9	.9	.9	1.0
.085	1.0	.9	.9	.9	1.0
.087	1.0	.9	.9	.9	1.0
.097	1.0	.9	.9	.9	1.0
.105	1.0	.9	.9	.9	1.0
.113	1.0	.9	.9	.9	.9
.120	.9	.8	.9	.8	.9

Table XXXVII

IO Time Ridge Regression Evaluation Statistics

Tapes	Cards	Lines	DiskIO	TapeIO	IO Time		Percent Error
					Actual	Calculated	
2	534	905	226	232	259.1235	270.113	4.24
0	28	7	28	0	1.9877	8.137	309.37
3	605	477	2505	385	553.9414	562.864	1.61
2	1072	2076	1963	299	454.9939	462.718	1.70
0	483	322	431	0	34.9514	46.459	32.92
1	885	8368	534	693	814.8247	760.480	- 6.67
0	63	10	300	0	22.3143	26.339	18.04
2	25	2275	2679	36	210.5129	250.663	19.07
2	13	1021	855	181	242.3307	253.986	4.81
3	107	21	31	807	809.4993	769.302	- 4.97
0	74	20	112	0	7.8376	12.629	86.65
2	207	634	304	323	357.2274	347.552	- 2.71
3	8	420	248	85	113.1524	135.126	19.42
2	54	917	185	31	50.8446	76.326	50.12

The percent error for this equation was not as close as expected. The percent error ranged from -6.67 to 309.37. The average relative error was 38.11 percent, with an average absolute error equaling 40.15 percent. Accepting the two largest values as outliers (they are the same cases considered as outliers for the turnaround model and exhibit the same characteristics here) did improve the accuracy. Now the mean relative error was 11.45 percent and the mean absolute error was 13.86 percent. Nevertheless, this model did not explain the expected amount of the variation of IO time.

V. Comparison of Techniques

Explained Criterion Variability

All three techniques explained similar percentages of the variability of the criterion variables. The following table contains the R-squared values achieved by each technique, indicating how well the data fit the various models.

Model	Turnaround Time	IO Time
Multilinear Regression	0.50910	0.98629
AID - Trimmed	0.47400	0.85300
AID - Complete	0.55600	0.93900
Ridge Regression	0.51710	0.96900

The AID - Trimmed figures denote the values at Level 3 where the tree was trimmed; the AID - Complete data reflect the variability explained by the total procedure, including all the leaves. No procedure adequately modeled turnaround time, thus indicating more data, currently unavailable through CPESIM, is required to adequately explain the variation. Therefore, the models produced would not be expected to have high predictive accuracy for turnaround time. However, almost all the variance in IO time was caused by measurable factors which were recorded in the CPESIM accounting log. The models produced for this variable should thereby be more accurate than those for turnaround time.

Results Available

Both multilinear and ridge regression techniques produce equations to

model the system. Neither technique, however, adequately modeled turnaround time. Both techniques indicated probable multicollinearity existed between variables IOtime and tapeIO. *Ridge regression* was designed to compensate for multicollinearity, and may therefore have been expected to produce a more accurate model for turnaround time. However, the unnormalized coefficient values for these variables equaled zero for all k greater than or equal to 0.024 and 0.015, respectively, while the system did not stabilize until k = 0.108. Thus these variables were not included in the model and ridge regression had no multicollinearity to compensate for in the actual turnaround time model equation. Further, for the IO time model, IOtime was a critical rather than predictor variable, and multicollinearity could therefore not exist in that model.

As expected, neither turnaround time equation produced accurate results. Table XXXVIII documents the error rates of the techniques. These high rates further confirmed the inadequacy of the models for this variable. The percentage of explained variation of IO time was quite high and satisfactory for all three models. However, neither equation produced by the two regression techniques was as accurate as expected. The error rates of both techniques are also summarized in Table XXXVIII. Nevertheless, the accuracy of these models was significantly greater than those modeling turnaround time.

Multilinear regression also produces information concerning the order in which the variables explained the variation and thus should be added to the equation. Further, the statistics provide guidance on how many variables should be added to the equation, and which add too little to the explanation of variation and may be excluded. Ridge regression produces little such guidance. Variables may be excluded from the equation if their normalized regression coefficients rapidly approach zero as the value of k increases.

Table XXXVIII

Accuracy Comparison of Multilinear and Ridge Regressions

Model	Turnaround Time Mean Errors		IO Time Mean Errors	
	Relative	Absolute	Relative	Absolute
Multilinear Regression	155.85%	182.60%	3.76%	12.59%
Multilinear Regression without outliers	76.79%	105.60%	- 2.84%	3.76%
Ridge Regression	283.06%	307.03%	38.11%	40.15%
Ridge Regression without outliers	53.77%	81.74%	11.45%	13.86%
Ridge Regression without tapes	n/a	n/a	29.17%	38.11%
Ridge Regression without tapes or outliers	n/a	n/a	1.03%	9.87%

However, ridge regression indicates neither which predictor variables account for sufficient variability to be added to the equation, nor which actually have little or no significant impact. Ridge regression does provide a visual record of the interaction of the predictor variables through the ridge trace.

AID produces a far different model. No specific equation is developed to account for current performance or predict future performance. Rather, the data are grouped into subsets which produce similar values for the criterion variables. Information produced here also includes how the groups are formed, the variation explained by each subgroup produced, and the standard deviation of the variance of the criterion variable within each group. No information is specifically produced on the values of the variables for those data records within each subgroup, except for those on which the data groups are split. The groupings identified can be studied further to obtain this information.

Ease of Use

The two types of regression processed were equally easy to use. Both accepted the data directly from the input file with no manipulation or conversion. Processing was straightforward. The multilinear regression procedure constructed a regression equation with only one run. Ridge regression produced an initial estimate of the optimal value of the bias factor, k, with the initial run. Subsequent runs converged on a more exact value.

In direct contrast, however, AID required significant data manipulation to group the data into manageable subsets. This action is required to insure the algorithm is viable to process large data files within a reasonable time. This grouping procedure is not difficult if each predictor variable is defined as being composed of equal length intervals. Here the

analyst must only determine the maximum and minimum values of each variable and determine the appropriate length to use to divide the total interval. However, the analyst has little control over how many data records are in each subgroup, or whether the divisions occur at natural boundaries, such as the actual groupings of the memory requirements in the CPESIM data. Therefore, the analyst may prefer to have greater control over the specific data records included in each subgroup. This would require a previous analysis of the data to determine where the divisions should occur. One method for this would be to use the SPSS Frequencies procedure, though other methods are also available. While this does require a significant amount of work to prepare the data for processing, the attributes of the system being studied may require such an analysis. Further, AID accepts only integer data, requiring the transformation of fixed or floating point data. Thus the analyst must also study each such variable to determine whether to convert the data through round-off, truncation, decimal point deletion, or value recoding. After all this preparation is completed, AID is simple to run. A significant amount of work is also required, however, after AID processing is complete to analyze and document all the vital information about the subgroups produced to determine what the model actually is, and what values of the criterion variable can be expected for predictor variables which meet the requirements of each subgroup. However, after this analysis, the model indicates that for given values (or ranges of values) of the predictor variables, the value of the criterion variable will also fall within a specific range for a significant percentage of the time. AID was the most difficult technique to use.

Algorithm Limitations

No restrictions or limitations exist with the SPSS multilinear regression procedure. However, the ridge regression program used in this study accepts only 16 total predictor and criterion variables. This limits its capabilities for real world data where many more variables may be involved in documenting system operation. Successive runs using different variables might be processed to eliminate variables where possible. Alternatively, another modeling technique might be first applied to determine the most significant variables, and ridge regression applied to those.

AID also has limitations. It accepts only integer data, thus requiring conversion of fixed or floating point data. Further, it too limits the number of predictor variables which may be supplied. However, since the limit is 80 such variables, the algorithm should easily handle most situations. A more serious problem is the lack of vital information about the subgroups identified by the method. Other statistical procedures must be run to provide more specific information, such as the range, mean, and standard deviation of each variable, including the criterion variable. These data are required before the system can be completely modeled.

CPESIM Summary

Multilinear regression produced the more accurate modeling equation for both variables. The bias factor involved in ridge regression adversely affected the accuracy of that equation since that model included no multicollinearity. AID provided the most useable information to provide a viable predictive model for turnaround time.

VI. Conclusions and Recommendations

This thesis effort involved modeling the CPESIM data base using three empirical modeling techniques: Multilinear regression, ridge regression, and automatic interaction detection (AID). All three techniques provide valuable information to model a computer system as part of a computer performance evaluation effort. Multilinear regression has long been the workhorse of empirical models for such projects, and its capabilities and limitations are well documented. Ridge regression was designed to a large extent to compensate for multicollinearity among the predictor variables. However, for the system under study, multicollinearity was not a factor in either model. Thus, this technique could not be expected to significantly improve the explanation of the criterion variability, and did not. One definite advantage of multilinear regression over ridge regression is its ordering of the predictor variables and the clear indication of which of them explain enough of the variability to be logically included in the equation. Nevertheless, for a system with a high degree of multicollinearity, ridge regression could well prove invaluable.

AID produces no explicit modeling equation, instead separating the data records into groups which produce similar ranges of values for the criterion variable. The groups specify the range into which the value of the criterion variable will fall for specific combinations of the predictor variables. This may not be sufficiently exact for systems which could be well modeled by multilinear regression. However, some systems and criterion variables, such as CPESIM and turnaround time, cannot be adequately explained through other techniques tested thus far. For such systems, AID could well prove most valuable. For this effort, multilinear regression only identified the

two variables which most directly influenced the value of turnaround time. AID identified the specific groupings of these variables which related directly to specific value ranges of the criterion variable. Thus, even if an exact equation cannot be constructed by some modeling technique to precisely predict the value of a criterion variable, AID can still provide an approximate range within which the value should fall.

Thus, multilinear regression remains a good, thorough method for modeling a system. Ridge regression should significantly improve the accuracy of the modeling equation for systems with much multicollinearity, although this was not proven in this effort. Finally, while AID can provide amplification for an already adequate model of the system, it may prove most valuable in providing basic organizing information about a system which cannot be adequately explained by another technique.

The techniques tested in this thesis effort should be applied to real world data with more variables and a more realistically large data base. This would better document the applicability of each technique to large-scale computer systems.

Other of the techniques identified for possible inclusion in this project should be examined using the same data base. Thus their results could be compared with those already derived to evaluate the relative performance and accuracy of the models.

Bibliography

1. Agrawala, A. K. and J. M. Mohr. "Some Results of the Clustering Approach to Workload Modelling," Proceedings of the Thirteenth Meeting of the Computer Performance Evaluation Users Group, 23-38, Washington D. C.: Computer Performance Evaluation Users Group, 1977.

2. Agrawala, A. K.; J. M. Mohr; and R. M. Bryant. "An Approach to the Workload Characterization Problem," Computer, 9: 18-29 (June 1976).

3. Chatterjee, Samprit and Bertram Price. Regression Analysis by Example. New York: John Wiley and Sons, 1977.

4. Cruise, D. R. Optimal Regression Models for Multivariate Analysis (Factor Analysis). China Lake, California: Naval Weapons Center, August 1977.

5. Ferrari, Dominico. Computer System Performance Evaluation. Englewood Cliffs, New Jersey: Prentice Hall, 1978.

6. Gooch, L. L. Policy Capturing with Local Models: The Application of the Automatic Interaction Detection Technique in Modeling Judgment. Unpublished PhD. Dissertation. University of Texas, Austin, Texas, 1972.

7. Gudes, Ehud and Charles Sechler. "Measures for Workload and Their Relation to Performance Evaluation," Proceedings of the Ninth, Tenth, and Eleventh Meetings of the Computer Performance Evaluation Users Group, 115-122. Washington D. C.: Computer Performance Evaluation Users Group, 1974-1975.

8. Hartrum, Thomas C. "Development of a Computer Turnaround Time Model," unpublished paper submitted for QBA 725, Wright State University, Dayton, Ohio, June 1979.

9. -----. Lecture Materials distributed in EE752, Computer Performance Evaluation, School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, 1980.

10. Hemmerle, W. J. "An Explicit Solution for Generalized Ridge Regression," Technometrics, 17: 309-314 (August 1975).

11. Hemmerle, W. J. and T. F. Brantle. "Explicit and Constrained Generalized Ridge Estimation," Technometrics, 20: 105-119 (May 1978).

12. Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32: 1-49 (March 1976).

13. Hoerl, Arthur E. and Robert W. Kennard. "Ridge Regression: Appli-

cations to Nonorthogonal Problems," Technometrics, 12: 69-82 (February 1970).

14. -----. "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12: 55-67, (February 1970).

15. Jain, Aridanan K. "Guideline to Statistical Approaches in Computer Performance Evaluation Studies," Performance Evaluation Review, 45-59. New York: Association of Computing Machinery Sigmetrics Spring-Summer 1979.

16. Kelly, John C. "The Statistical Nature of Computer Performance Evaluation," Proceedings of the Ninth, Tenth, and Eleventh Meetings of the Computer Performance Evaluation Users Group. 5-18. New York: Computer Performance Evaluation Users Group, 1974-1975.

17. Kobayaski, Hisaki. Modeling and Analysis: An Introduction to System Performance Evaluation Methodology. Reading, Massachusetts: Addison-Wesley Publishing Co., 1978.

18. Lewis, P. A. W. and G. S. Shedler. Analysis and Modelling of Point Processes in Computer Systems. MS Thesis. Naval Postgraduate School, Monterey, California, September 1977. (AD A050247).

19. Little, Roderick J. A. "Consistent Regression Methods for Discriminant Analysis with Incomplete Data," American Statistical Association Journal, 329-335 (June 1978).

20. Marquardt, Donald W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimators," Technometrics, 12: 591-612 (August 1970).

21. McNichols, Charles W. An Introduction to: Applied Multivariate Data Analysis, unpublished text. School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, 1980.

22. Morton, T. Gregory. "Factor Analysis, Multicollinearity, and Regression Appraisal Models," Appraisal Journal, 578-592 (October 1977).

23. Nie, Norman H.; C. Hadlai Hull; Jean G. Jenkins; Karin Steinbrenner; and Dale H. Bent. Statistical Package for the Social Sciences (Second Edition). New York: McGraw-Hill Book Co., 1975.

24. Nutt, Gary J. Computer Systems Monitoring Technique. Springfield, Virginia: National Science Foundation, February 1977.

25. Obenchain, R. L. "Classical F-Tests and Confidence Regions for Ridge Regression," Technometrics, 19: 429-440 (November 1977).

26. "Preliminary Description: RIDGE, A Program to Support Ridge Regression Analysis," notes, Department of Operational Sciences. School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB.

27. Schroeder, Anne. "How Multidimensional Data Analysis Techniques Can be of Help in the Study of Computer Systems," Proceedings of the Fourteenth Meeting of the Computer Performance Evaluation Users Group. 149-166. Washington D. C.: Computer Performance Evaluation Users Group, 1978.

28. Strawderman, William E. "Minimal Adaptive Generalized Ridge Regression Estimators," American Statistical Association Journal, 623-628 (September 1978).

29. Thompson, Jimmy W. and Thomas C. Hartrum. "The Application of Clustering Techniques to Computer Performance Modeling," Proceedings of the Fifteenth Meeting of the Computer Performance Evaluation Users Group. 147-162. Washington D. C.: Computer Performance Evaluation Users Group, 1979.

30. Watson, R. Computer Performance Analysis: Applications of Accounting Data. Santa Monica, California: Rand Corp., 1971.

VITA

Margaret A. Stover was born on 13 January 1953 in Little Rock, Arkansas. She graduated from high school in Little Rock, Arkansas in 1970 and attended Hendrix College, Conway, Arkansas, from which she received the degree of Bachelor of Arts in Mathematics in June 1974. Upon graduation she studied Computer Science at the University of Arkansas at Little Rock, Little Rock, Arkansas from which she received no degree. Concurrently she was employed by the University of Arkansas Medical Center. She entered the Air Force on active duty in June 1975, and received her commission from Officer Training School in August 1975. She was assigned to the 1139 Comptroller Services Squadron at Bolling AFB, later renamed Det 1, 1500 Computer Services Squadron (MAC Washington Area Computer Center) and located at Andrews AFB, where she served as systems analyst, operations officer, and executive support officer until entering the School of Engineering, Air Force Institute of Technology, in June 1980. She was named MAC Data Automation Officer of the Year for 1979.

Permanent address: 1005 N. Taylor St.
Little Rock, Arkansas 72205

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/GCS/EE/81D-17	2. GOVT ACCESSION NO. AD A115	3. RECIPIENT'S CATALOG NUMBER 550
4. TITLE (and Subtitle) APPLICATION OF STATISTICAL ANALYSIS TECHNIQUES TO COMPUTER PERFORMANCE EVALUATION		5. TYPE OF REPORT & PERIOD COVERED MS Thesis 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Margaret A. Stover Capt		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Air Force Institute of Technology (AFIT-EN) Wright-Patterson AFB, Ohio 45433		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Det 1, 1500 Computer Services Squadron (MAC Washington Area Computer Center) Andrews AFB, Maryland 20331		12. REPORT DATE December 1981 13. NUMBER OF PAGES 116
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

15 APR 1982

18. SUPPLEMENTARY NOTES

Approved for public release IAW AFR 190-17

FREDERIC C. LYNCH, Major, USAF
Director of Public Affairs

Dean for Research and Professional Development
Air Force Institute of Technology (ATC)
Wright-Patterson AFB, OH 45433

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multivariate Statisticas Analysis
Computer Performance Evaluation
Multilinear Regression
Ridge Regression
Automatic Interaction Detection

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Multivariate statistical analysis techniques, multilinear regression, automatic interaction detection, and ridge regression, were applied to computer performance evaluation. A simulated data base was modeled using each technique. Multilinear regression derived the most accurate modeling equations, though no method adequately explained the variation in turnaround time. Ridge regression, designed to circumvent multicollinearity, also created modeling equations, but was less accurate because no multicollinearity was present in the system. AID clustered the data into groups which explained high levels of the variability of

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

the criterion variable within each group. Data records which met the selection criteria for each group should also have values of the criterion variable within the range specified for/by that group. Although difficult to use, AID provided initial information about systems which could not be modeled through regression or other techniques. The applicability of each technique to selected types of computer systems was also documented.